Bayesian P-spline mixture modeling of nonstationary extreme forest temperatures

B. Oumow, M. B. de Carvalho, and A. C. Davison
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Abstract. Extreme forest temperatures can be responsible for huge ecological damages, such as increasing the risk of wildfires or leading to severe modifications in vegetation. In this paper we propose to model the conditional distribution of forest temperatures using a Bayesian P-spline mixture method for nonstationary extremes. Our approach entails modeling the bulk of the distribution and the parameters of an asymptotically-motivated model for the tails, through Bayesian semiparametric generalized additive models using B-splines and penalties. We apply our method to study how extreme forest temperatures in Switzerland have been evolving over the last decade.

Keywords: Bayesian P-spline, Forest ecosystems, Generalized additive model, Nonstationary extremes, Extreme value mixture model, Statistics of extremes, Temperature data

1. Introduction

Extreme-value mixture models are an increasingly popular class of methods for modeling simultaneously the bulk of a distribution and its tails. They have been proved to be highly flexible for modeling data, and several procedures have been recently proposed in the literature. In some approaches a parametric form is assumed for the bulk of the distribution (Frigressi *et al.*, 2002; Mendes and Lopes, 2004; Carreau and Bengio, 2009), while other authors suggest estimating the distribution of the bulk with semiparametric and nonparametric techniques (Tancredi *et al.*, 2006; Macdonald *et al.*, 2011). None of these authors has treated the case of nonstationary processes—which is of central interest in extreme-value modeling—, and it is our aim here to propose a model for this framework.

Our work is motivated by the catastrophic impact of extreme temperatures on forest ecosystems. Recent studies have shown that certain species of trees are disappearing from Swiss forests due to climate change. For a long time, biologists have mainly taken an interest in the progression of average temperatures, but it is now acknowledged that it is mostly extreme climate situations that actually modify vegetation, as certain species cannot survive to extremely low or high temperatures. For example, the number of days when the average temperature has exceeded 20 °C has doubled in Viège over the last 20 years, causing in this area the death of half of the scots pines, a tree that cannot stand too high temperatures. A greater understanding of the dynamics governing how extreme temperatures evolve in forest microclimates is thus of large ecological interest. We are prompted by temperature data from 14 forest weather stations in Switzerland whose locations are represented in Figure 1, and which cover different biogeographical zones of Switzerland, with each forest being characterized by some vegetation and environmental features including altitude, orientation, soil type, and dominant tree species (Ferrez *et al.*, 2011). All these series are nonstationary, thus precluding the application of any of the above-mentioned extreme-value mixture models.

Address for correspondence: A. C. Davison, Institute of Mathematics, Analysis, and Applications; Ecole Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland.
E-mail: Anthony.Davison@epfl.ch

Figure 1: Locations of the 14 forest weather stations for which data are available.

When nonstationary extremes need to be modeled, nonparametric regression and smoothing methods are usually employed. The extension of these methods to extreme-value modeling involves computational difficulties especially due to the nonstandard form of the likelihood. Local likelihood models have been developed by Davison and Ramesh (2000), Hall and Tajvidi (2000), and Ramesh and Davison (2002). Chavez-Demoulin and Davison (2005) used generalized additive models with penalized likelihood estimation methods to smooth sample extremes. A similar modeling strategy is taken by Padoan and Wand (2008) and Laurini and Pauli (2009) who consider generalized linear mixed models with penalized splines.

In this paper we propose an extreme-value mixture model for nonstationary extremes using Bayesian P-spline generalized additive models (Lang and Brezger, 2004; Brezger and Steiner, 2008). Our method allows us to model dependence of the bulk and the tails on multiple covariates with smooth functions, thus being more flexible than existing models. Our mixture model includes the threshold as a parameter to be estimated, thus overcoming the problems of threshold selection and uncertainties in the inference process. This is an important feature of extreme-value mixture models compared to other existing solutions for threshold choice (Northrop and Jonathan, 2011).

Section 2 provides an overview of relevant aspects of statistics of extremes and extreme-value mixture models, and Section 3 discusses their extension to nonstationary data using P-spline modeling. Markov chain Monte Carlo (MCMC) inference is discussed in Section 4, whereas Section 5 provides some numerical experiments to demonstrate the performance of the model and estimation procedure in a particular case. Finally, Section 6 describes the application to extreme forest temperatures. A brief discussion concludes the paper.

2. Background on stationary extremes

2.1. Point process representation

Let $Y_1, \ldots, Y_n$ be a stationary sequence of independent and identically distributed random variables. For suitable sequences of normalising constants $\{a_n > 0\}$ and $\{b_n\}$, a sufficiently large u, the point process $P_n = \{(i/(n+1), Y_{n,i}) : Y_{n,i} = (Y_i - b_n)/a_n,\ i = 1, \ldots, n\}$

converges on regions of the form $(0,1)\times[u,\infty)$ to a time-homogeneous Poisson process, with intensity measure

$$\Lambda(\,[t_1,t_2]\times[z,\infty)\,) = \begin{cases} (t_2-t_1)\{1+\xi(z-\mu)/\sigma\}_+^{-1/\xi}, & \xi\neq 0,\\ (t_2-t_1)\exp\{-(z-\mu)/\sigma\}, & \xi=0, \end{cases} \tag{1}$$

defined on $\{z\in\mathbb{R}: 1+\xi(z-\mu)/\sigma>0\}$ with $\sigma>0$, where $(\,\cdot\,)_+$ is the positive-part function; the parameters μ, σ, and ξ are referred to as the location, scale, and shape parameters. To overcome problems of definition in the inference, it is useful to consider the intensity parameter $\lambda=\{1+\xi(u-\mu)/\sigma\}^{-1/\xi}$, for a given threshold u. The model is then equivalently characterized by the parameters λ, σ, and ξ. The conditional sizes of the k exceedances over u, $W_{k,1}=Y_{k,1}-u,\ldots,W_{k,k}=Y_{k,k}-u$, are distributed according to a generalized Pareto distribution

$$\mathrm{GPD}(w) = \begin{cases} 1-(1+\xi w/\psi_u)_+^{-1/\xi}, & \xi\neq 0,\\ 1-\exp(-w/\psi_u), & \xi=0, \end{cases} \tag{2}$$

where ψ_u is related to the Poisson process parameters by $\psi_u=\sigma+\xi(u-\mu)=\sigma\lambda^{-\xi}$. These models are still valid for stationary series that satisfy regularity conditions on long- and short-range dependence of the extremes.

2.2. Extreme-value mixture model

Suppose there exists a pair of unknown thresholds $u=(u_{\mathrm{L}},u_{\mathrm{R}})$, with $u_{\mathrm{L}}<u_{\mathrm{R}}$, that allows us to group the data into blocks corresponding to the tails and the bulk of the distribution of Y. The data corresponding to the left tail, bulk, and right tail are respectively defined by the random sets $\mathcal{L}=\{i: y_i<u_{\mathrm{L}}\}$, $\mathcal{B}=\{i: u_{\mathrm{L}}\le y_i\le u_{\mathrm{R}}\}$, and $\mathcal{R}=\{i: y_i>u_{\mathrm{R}}\}$. We use the letters corresponding to these random sets as subscripts or superscripts to identify if either the data or the parameters are from the tails or the bulk; for example X_{B} denotes the covariates of the design matrix X corresponding to the bulk, and $\xi^{(\mathrm{R})}$ represents the shape parameter of the generalized Pareto distribution for the right tail. Unless otherwise stated, we use 'S' as a subscript or superscript to denote an expression which holds for both tails, so that for example when we write $\xi^{(\mathrm{S})}=0.5$ it should be understood that $\xi^{(\mathrm{L})}=\xi^{(\mathrm{R})}=0.5$; in some exceptions 'S' will also be used to consider the tails and the bulk, but we will make that explicit in such cases. The extreme-value mixture model can be expressed as a mixture model

$$\varphi(y) = p_{\mathrm{L}}\varphi_{\mathrm{L}}(y) + p_{\mathrm{B}}\varphi_{\mathrm{B}}(y) + p_{\mathrm{R}}\varphi_{\mathrm{R}}(y), \tag{3}$$

where p_{L} and p_{R} are respectively the probabilities of being below u_{L} and above u_{R}, to scale the relative contributions represented by bulk and tail components Macdonald *et al.* (2011). In practice these are estimated using the proportion of data above the threshold, that is $\hat{p}_{\mathrm{L}}=\hat{\lambda}^{(\mathrm{L})}/n$ and $\hat{p}_{\mathrm{R}}=\hat{\lambda}^{(\mathrm{L})}/n$. Moreover,

$$\begin{aligned} \varphi_{\mathrm{L}}(y) &= \mathrm{gpd}(-y\mid -u_{\mathrm{L}},\psi_{u_{\mathrm{L}}},\xi^{(\mathrm{L})})I_{(-\infty,u_{\mathrm{L}})}(y),\\ \varphi_{\mathrm{B}}(y) &= \frac{h(y)}{H(u_{\mathrm{R}})-H(u_{\mathrm{L}})}I_{[u_{\mathrm{L}},u_{\mathrm{R}}]}(y),\\ \varphi_{\mathrm{R}}(y) &= \mathrm{gpd}(y\mid u_{\mathrm{R}},\psi_{u_{\mathrm{R}}},\xi^{(\mathrm{R})})I_{(u_{\mathrm{R}},\infty)}(y), \end{aligned} \tag{4}$$

where gpd denotes the density of the generalised Pareto distribution—which is tantamount to a Poisson–GPD approach with parametrization $(\lambda^{(\mathrm{S})},\psi^{(\mathrm{S})},\xi^{(\mathrm{S})})$ in §2.1, where $\psi^{(\mathrm{S})}=\psi_{u_{\mathrm{S}}}$ for S = L or R.

Model (3) may be inadequate for various reasons. The first one is that Poisson process limit (1) holds for infinite datasets. In practice, available quantity of data is finite but the Poisson–GPD model often fits well. Another problem arises from the fact that regularity conditions on short-range dependence may be unsatisfied, showing clusters of extremes. This can be solved by retaining only maxima of independent clusters (Leadbetter *et al.*, 1983), but a declustering algorithm is necessary and some information is lost. This problem is of secondary importance if we focus on long-term trends.

3. Bayesian P-spline mixture model for nonstationary extremes

3.1. Local model

The above discussion assume that the underlying process is stationary. We would like to extend the extreme-value mixture model discussed in §2.2 to nonstationary data. The nonstationarity for the extremes values arises for example when threshold exceedances becomes more or less frequent with time or when the size of the exceedances changes over time. Suppose that our data are a response variable $Y \in \mathbb{R}^{n\times 1}$, a design matrix $X \in \mathbb{R}^{n\times p}$, and a matrix of further covariates $V \in \mathbb{R}^{n\times q}$. The starting point for our modeling is a conditional version of (3) which is given by

$$\varphi(y \mid x, v) = p_{\mathrm{L}}(x, v)\varphi_{\mathrm{L}}(y \mid x, v) + p_{\mathrm{B}}(x, v)\varphi_{\mathrm{B}}(y \mid x, v) + p_{\mathrm{R}}(x, v)\varphi_{\mathrm{R}}(y \mid x, v), \tag{5}$$

where $p_{\mathrm{S}}(x, v)$ and $\varphi_{\mathrm{S}}(y \mid x, v)$ are defined in a similar way as in §2.2 following this conditional representation.

We assume that the conditional responses in the bulk follow an exponential family distribution

$$h(y_i \mid v_i, x_i) = c(y_i, \phi) \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} \right\}, \quad i \in \mathcal{B},$$

where $b(\cdot)$, $c(\cdot)$, θ_i, and ϕ define the respective distribution, and we model these observations using the generalized additive model (Hastie and Tibshirani, 1990)

$$\eta(v_i, x_i) = v_i^{\mathrm{T}} d + \sum_{j=1}^{p} l_j(x_{ij}) = g(\,\mathbb{E}[\,y_i \mid v_i, x_i\,]\,), \quad i \in \mathcal{B}. \tag{6}$$

Here $g(\cdot)$ is a link function, the $l_j(\cdot)$ are arbitrary smooth univariate functions, and d is a parameter vector; to ease notation we use η_i to denote $\eta(v_i, x_i)$. The definition of bulk needs to be updated here since we are now concerned with conditional distributions. The thresholds are now defined in a way that their distance to the linear predictor of the bulk η is given by a random quantity, i.e.

$$u_{\mathrm{S}}(v, x) = \eta(v, x) + c_{\mathrm{S}}, \tag{7}$$

with $c_{\mathrm{L}} < 0$ and $c_{\mathrm{R}} > 0$. The bulk and the tails are thus defined in a similar way to §2.2, but we replace the y_i with η_i, and take the covariate-dependent thresholds (7) into account. The specification in (7) can seem relatively strong but is reasonable to model data with homoscedasticity in the bulk and facilitates the fit. This specification can be relaxed by letting $c_{\mathrm{S}}(v, x)$ have some parametric form, e.g. a straight line, a polynomial, or nonlinear transformations of these. Numerical experiments in §5 show one case where a more flexible model is necessary.

To model the left and right tails we consider the point process approach locally by supposing that the Poisson process parameters $\Theta_{\mathrm{S}} = (\lambda^{(\mathrm{S})}, \psi^{(\mathrm{S})}, \xi^{(\mathrm{S})})^{\mathrm{T}}$ also depend on covariates. We consider a semiparametric generalized additive model for each of the tails

$$\Theta_{\mathrm{S}}(v,x) = \left\{ \begin{array}{c} \exp\left(v^{\mathrm{T}}a^{(\mathrm{S})} + \sum_{j=1}^{p} f_j^{(\mathrm{S})}(x_j)\right) \\ \exp\left(v^{\mathrm{T}}b^{(\mathrm{S})} + \sum_{j=1}^{p} g_j^{(\mathrm{S})}(x_j)\right) \\ v^{\mathrm{T}}c^{(\mathrm{S})} + \sum_{j=1}^{p} h_j^{(\mathrm{S})}(x_j) \end{array} \right\}, \tag{8}$$

where x and v are are respectively p and q-vectors of covariates corresponding to the design matrices X and V; we store the regression vectors and the smooth functions in (8) in the matrices $\mathbb{D}^{(S)} = (a^{(\mathrm{S})}, b^{(\mathrm{S})}, c^{(\mathrm{S})})^{\mathrm{T}}$ and $\mathbb{L}_j^{(\mathrm{S})} = (f_j^{(\mathrm{S})}, g_j^{(\mathrm{S})}, h_j^{(\mathrm{S})})^{\mathrm{T}}$, respectively. For simplicity we assume that all the above-mentioned additive models are based on the same covariates v and x, but there is no difficulty in expanding this to a more general case. Following Chavez-Demoulin and Davison (2005), we divide the interval $[0,1]$ into m subintervals with a sufficiently small length $\delta > 0$, such that $\lambda(t)$ is constant over each interval, so that the Poisson–GPD likelihood can be approximated by

$$\mathscr{L}_{\mathcal{S}}(\Theta_{\mathrm{S}}, \eta, c_{\mathrm{S}} \mid \cdot) \propto \exp\left\{-\delta \sum_{k=1}^{m} \lambda_{(n+1)k\delta}^{(\mathrm{S})}\right\} \prod_{i\in\mathcal{S}} \frac{\lambda_i^{(\mathrm{S})}}{\psi_i^{(\mathrm{S})}} \times \left[1 + \xi_i^{(\mathrm{S})} \frac{\mid y_i - (\eta_i + c_{\mathrm{S}}) \mid}{\psi_i^{(\mathrm{S})}}\right]^{-1/\xi_i^{(\mathrm{S})}-1}, \tag{9}$$

where $\mathrm{S} = \mathrm{L}$, for $\mathcal{S} = \mathcal{L}$, and $\mathrm{S} = \mathrm{R}$, for $\mathcal{S} = \mathcal{R}$. Here $\lambda_i^{(\mathrm{S})}$, $\psi_i^{(\mathrm{S})}$, $\xi_i^{(\mathrm{S})}$ are the realisations of the corresponding parameters at (v_i, x_i), and have to be replaced by their additive expressions, for $i = 1, \ldots, n$. We remark that the likelihood (9) factorizes into two orthogonal components, one involving $\lambda^{(\mathrm{S})}$ is the likelihood for the number of exceedances N, hence characterising the tail probabilities p_{S} in the mixture model (5), while the other, involving $\psi^{(\mathrm{S})}$ and $\xi^{(\mathrm{S})}$, is the likelihood for the size of the exceedances conditional on N corresponding to the GPD densities in (5). Inferences on $\lambda^{(\mathrm{S})}$ can therefore be separated from those on $(\psi^{(\mathrm{S})}, \xi^{(\mathrm{S})})$. Finally, the likelihood for the bulk is

$$\mathscr{L}_{\mathcal{B}}(\eta, c_{\mathrm{L}}, c_{\mathrm{R}} \mid Y, X, V) \propto \prod_{i\in\mathcal{B}} \frac{c(y_i, \phi) \exp\{y_i\theta_i - b(\theta_i)\}/\phi}{H(\eta_i + c_{\mathrm{R}}) - H(\eta_i + c_{\mathrm{L}})}. \tag{10}$$

The likelihood of our extreme-value mixture model can be separated into the contributions from the bulk (10) and the tails (9)

$$\mathscr{L}(\Theta \mid Y, X, V) = \mathscr{L}_{\mathcal{L}}(\Theta_{\mathrm{L}}, \eta, c_{\mathrm{L}} \mid \cdot)\mathscr{L}_{\mathcal{B}}(\Theta_{\mathrm{B}}, c_{\mathrm{L}}, c_{\mathrm{R}} \mid \cdot)\mathscr{L}_{\mathcal{R}}(\Theta_{\mathrm{R}}, \eta, c_{\mathrm{R}} \mid \cdot), \tag{11}$$

where Θ is the vector of all the parameters to be estimated in the overall model and Θ_{B} is the vector of the parameters involved in the additive model for the bulk.

3.2. Bayesian P-spline additive modeling

We propose to approximate the unknown functions $\mathbb{L}_j^{(\mathrm{S})}$ and l_j using a spline of degree s with equally spaced knots $x_{j,\min} = k_{j1} < \cdots < k_{j\kappa} = x_{j,\max}$, for $j = 1, \ldots, p$. This spline can be written in terms of a linear combination of $\kappa + s - 1 = \Psi$ B-spline basis functions B_{jm}, and to simplify notation we assume the same number of knots Ψ for every function. The overall model can now be written with the aid of the design matrices $Z_j^{(\mathrm{S})}(i,m) = B_{jm}(x_{ij})$, $\mathrm{S} \in \{\mathrm{L}, \mathrm{B}, \mathrm{R}\}$, respectively defined on $\mathbb{R}^{|\mathcal{S}|\times\Psi}$, $\mathcal{S} \in \{\mathcal{L}, \mathcal{B}, \mathcal{R}\}$, so that

$$\widetilde{\Theta}_{\mathrm{S}} = \left\{ \begin{array}{c} \exp\left(V_{\mathrm{S}}a^{(\mathrm{S})} + \sum_{j=1}^{p} Z_j^{(\mathrm{S})}\alpha_j^{(\mathrm{S})}\right) \\ \exp\left(V_{\mathrm{S}}b^{(\mathrm{S})} + \sum_{j=1}^{p} Z_j^{(\mathrm{S})}\beta_j^{(\mathrm{S})}\right) \\ V_{\mathrm{S}}c^{(\mathrm{S})} + \sum_{j=1}^{p} Z_i^{(\mathrm{S})}\gamma_j^{(\mathrm{S})} \end{array} \right\}, \quad \widetilde{\eta} = V_{\mathrm{B}}d + \sum_{j=1}^{p} Z_j^{(\mathrm{B})}\iota_j. \tag{12}$$

Here $\ell_j^{(\mathrm{S})} = (\alpha_j^{(\mathrm{S})}, \beta_j^{(\mathrm{S})}, \gamma_j^{(\mathrm{S})})$, and ι_j, are the coefficients of the jth predictor function on the B-spline basis for the corresponding parameters, while $\mathbb{D}^{(\mathrm{S})} = (a^{(\mathrm{S})}, b^{(\mathrm{S})}, c^{(\mathrm{S})})$ and d are the ones corresponding to fixed effects.

To ensure sufficient flexibility, we follow Eilers and Marx (1996) and use between 20 and 40 knots, but to avoid overfitting we use a roughness penalty on the kth-order differences of adjacent regression coefficients. To setup the priors for our model we define $\theta^{(j)} = (\ell_j^{(\mathrm{L})}, \iota_k, \ell_j^{(\mathrm{R})})^{\mathrm{T}} \in \mathbb{R}^{7\times\Psi}$, and $\Lambda = (\mathbb{D}^{(\mathrm{L})}, d, \mathbb{D}^{(\mathrm{R})})^{\mathrm{T}} \in \mathbb{R}^{7\times q}$. We introduce priors for the fixed effects parameters Λ_l and the coefficients $\theta_l^{(j)}$, for $l = 1, \ldots, 7$, and $j = 1, \ldots, p$; for the fixed effects parameters, and for the k first components of each $\theta_l^{(j)}$, we assume diffuse (constant) priors. Moreover, kth-order difference penalties are replaced by their stochastic analogues, that is, kth-order Gaussian random walks $D_k\theta_l^{(j)} \sim \mathrm{N}(0, \tau_{jl}^2 I_{\Psi-k})$, where $D_k \in \mathbb{R}^{(\Psi-k)\times\Psi}$ is the kth-order difference matrix. The amount of smoothness is controlled by the variance parameters τ_{jl}^2, which corresponds to the inverse of the smoothing parameter in the frequentist approach. The priors have the following global form

$$\pi(\theta_l^{(j)} \mid \tau_{jl}^2) \propto \left(\frac{1}{\tau_{jl}^2}\right)^{(\Psi-k)/2} \exp\left(-\frac{1}{2\tau_{jl}^2}(\theta_l^{(j)})^{\mathrm{T}} D_k^{\mathrm{T}} D_k \theta_l^{(j)}\right). \tag{13}$$

Hyperpriors are assigned to the variances τ_{jl}^2 in a hierarchical model by highly dispersed but proper inverse Gamma priors, that is

$$\tau_{jl}^2 \sim \mathrm{IG}(\lambda_{jl}, \nu_{jl}),$$

with hyperparameters λ_{jl} and ν_{jl}. Usually, we choose $\lambda_{jl} = \nu_{jl}$ to obtain almost diffuse priors, for example $\lambda_{jl} = \nu_{jl} = 0.001$; another common choice is to set $\lambda_{jl} = 1$ and ν_{jl} small, for example $\nu_{jl} = 0.005$ or $\nu_{jl} = 0.0005$. If we assume that the bulk distribution is a Gaussian with mean η and variance Σ^2, we also assign an inverse Gamma prior to Σ^2:

$$\Sigma^2 \sim \mathrm{IG}(\lambda_\Sigma, \nu_\Sigma).$$

Finally, we assume diffuse priors for the constant threshold parameters c_{L} and c_{R}, that is $c_{\mathrm{S}} \propto 1$.

4. MCMC inference

Let Θ be the vector of all parameters to be estimated in the global model. Bayesian inference is based on the full posterior distribution

$$p(\Theta \mid \cdot) \propto \mathscr{L}(\Theta \mid \cdot) \times \prod_{j=1}^{p}\prod_{l=1}^{7} \pi(\theta_l^{(j)} \mid \tau_{jl}^2)\pi(\tau_{jl}^2) \times \pi(c_{\mathrm{L}}) \times \pi(c_{\mathrm{R}}) \times \pi(\phi), \tag{14}$$

where $\mathscr{L}(\Theta \mid \cdot)$ is the overall likelihood of the model and can be separated in three parts as in (11), and Θ is now the vector of all the parameters to be estimated in the overall model. In (14) we avoided the contributions of the fixed effects parameters as their priors were assumed to be diffused. Because (14) is often analytically intractable, we employ MCMC techniques to obtain estimates for the parameters of interest. From this expression we can obtain the posterior distribution for each parameter.

The posterior distributions of $\theta_l^{(j)}$ and Λ_l, $l = 1, \ldots, 6$, which are respectively the spline and fixed effects coefficients for the tail parameters $\lambda^{(\mathrm{S})}$, $\psi^{(\mathrm{S})}$, $\xi^{(\mathrm{S})}$, are

$$p(\theta_l^{(j)} \mid \cdot) \propto \mathscr{L}_{\mathcal{S}}(\Theta_{\mathrm{S}} \mid \cdot)\pi(\theta_l^{(j)}), \quad p(\Lambda_l \mid \cdot) \propto \mathscr{L}_{\mathcal{S}}(\Theta_{\mathrm{S}} \mid \cdot),$$

where S = L if $l = 1, 3, 5$, S = R if $l = 2, 4, 6$, and the posterior distributions for the bulk parameters $\theta_7^{(j)}$ and Λ_7 are

$$p(\theta_7^{(j)} \mid \cdot) \propto \mathscr{L}(\Theta \mid \cdot)\pi(\theta_7^{(j)}), \quad p(\Lambda_l \mid \cdot) \propto \mathscr{L}(\Theta \mid \cdot).$$

As the priors for the threshold parameters c_{S}, S = L or R, are assumed to be diffuse, their posterior distribution is

$$p(c_{\mathrm{S}} \mid \cdot) \propto \mathscr{L}_{\mathcal{B}}(\Theta_{\mathrm{B}} \mid \cdot)\mathscr{L}_{\mathcal{S}}(\Theta_{\mathrm{S}} \mid \cdot).$$

Finally, the posterior distribution of the variance parameters τ_{jl}, $l = 1, \ldots, 7$, is

$$p(\tau_{jl} \mid \cdot) \propto \pi(\theta_l^{(j)} \mid \tau_{jl}^2)\pi(\tau_{jl}^2).$$

Following Behrens *et al.* (2004) and Macdonald *et al.* (2011), we use a blockwise Metropolis–Hastings algorithm to obtain samples of all the parameters. We have selected proposal distributions compatible with the restrictions of the model parameters, and we use random walk proposals. Computational details are given in the Appendix.

5. Numerical experiments

5.1. Data-generating processes

Numerical experiments have been made to demonstrate the performance of the model and estimation procedure. A wide range of scenarios was considered to study the behavior of the posterior distribution in different situations. The assumptions are the following. The convergence in the tails depends on the number of available observations, and hence we considered situations in which we have a relatively large number of observations to estimate the GPD parameters. We focus on cases where $n = 4000, 10000$ or 20000 and $\lambda^{(\mathrm{S})}$ is sufficiently large. We also focus on the particular case where a single covariate is considered with no fixed effects, and where the underlying smooth functions for the bulk of the distribution and for the tail parameters were chosen with low or moderate curvature. We also considered a case with a more complex threshold by letting c_{S} to have a quadratic form. The simulated observations corresponding to these two cases for $n = 4000$ are shown on Figure 3; the two thresholds are represented by the solid lines. The numerical experiments were performed as follows:

(a) Choose the (single) covariate t which consists in an equidistant grid of n design points between -3 and 3.

(b) Generate n bulk observations from $\eta_i = \mathrm{TN}(f(t_i), 1, u_{\mathrm{L}}(t_i), u_{\mathrm{R}}(t_i))$, where TN represent the truncated normal distribution, $u_{\mathrm{S}}(t_i) = f(t_i) \pm c_{\mathrm{S}}(t_i)$ are the thresholds, and the function f is sinusoidal: $f(t) = 1/0.72 \sin(t)$. A first case corresponds to the situation where $c_{\mathrm{S}} \equiv 2$; a second case to the situation where $c_{\mathrm{S}}(t) = 1.8 + 0.1t + 0.05t^2$.

(c) Allow each tail parameter to depend on the same covariate t as follows:

$$\Theta_{\mathrm{S}}(t) = \begin{pmatrix} 0.05n \exp\{\cos(t)\} \\ \exp\{\sin(t)/4\} \\ t^2/20 - 0.25 \end{pmatrix}.$$

(d) Generate n exceedances above each threshold from $\mathrm{GPD}(\psi_u^{(\mathrm{S})}, \xi^{(\mathrm{S})})$.

(e) For all t_i, we have therefore an observation in the bulk and one in each tail. Then for all $i = 1, \ldots, n$, we choose with probability $p_{\mathrm{S}} = \lambda_{\mathrm{S}}/n$, S $\in \{\mathrm{L}, \mathrm{B}, \mathrm{R}\}$, which one is conserved for the simulation. The observation vector of size n is denoted by y.

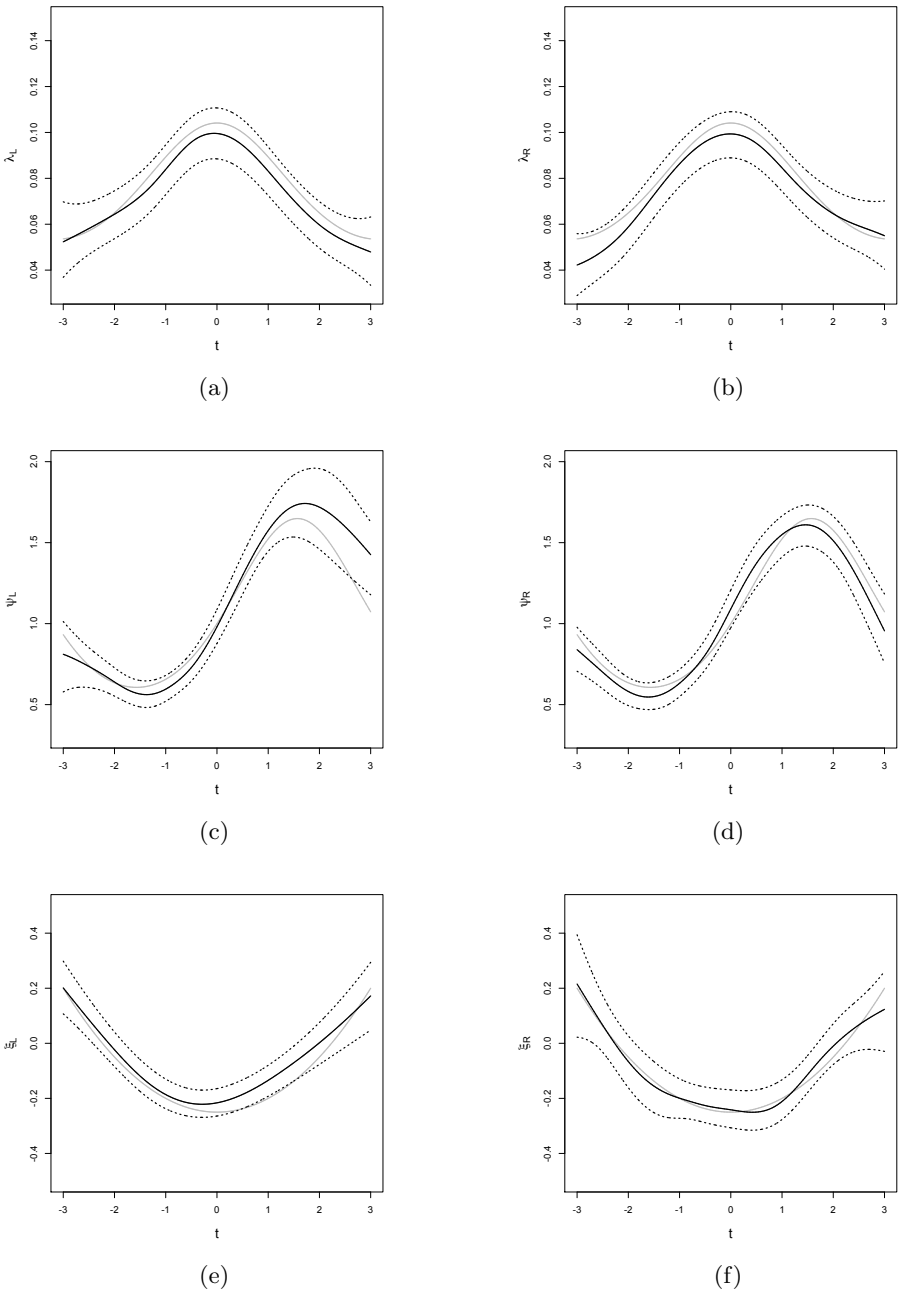
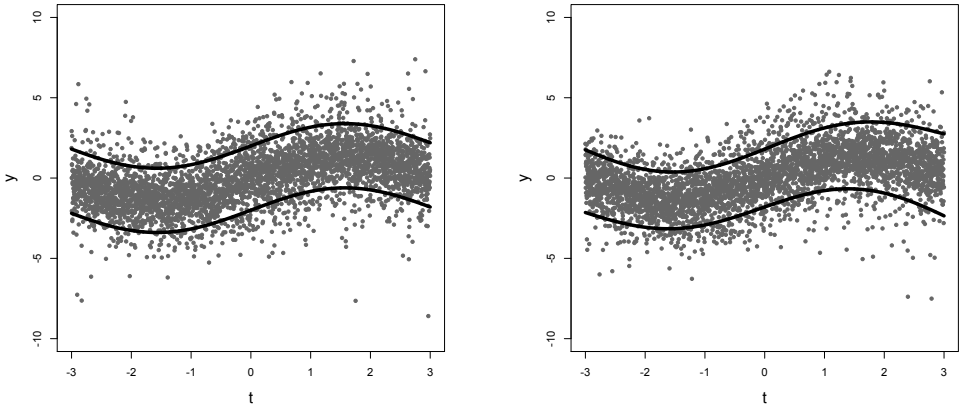


Figure 2: Smooth estimates of the tails parameters for the data in Figure 3 (a). The solid gray line represents the simulated function and the solid black line represents its estimate using our Bayesian P-spline mixture model with 15 knots and second order difference penalties; the dotted lines represent the 95% credible bands for the estimated functions.



(a) (b)

Figure 3: Simulated observations for $n = 4000$. The solid lines represent the thresholds: (a) $c_{\mathrm{S}} \equiv 2$; (b) $c_{\mathrm{S}}(t) = 1.8 + 0.1t + 0.05t^2$.

5.2. Results

The results of our numerical experiments for the case where $n = 4000$ are presented in Figures 4 and 2. We have chosen $k = 15$ knots, which corresponds to 17 B-spline basis functions. Our estimate of the bulk is suitable, with the smooth estimate being close to the true function; this is reported in Supporting Information. The estimate of c_{S} are also satisfying: we obtain respectively 2.02 and 2.01 for the left and right tails, with standard deviation of 0.0216 and 0.0170. The estimates of the Poisson–GPD parameters also yield good fits, with the general form of the corresponding smooth functions being recovered and their scaling being the correct one. We observe that the estimation is less precise at the boundaries, but this is due to the fact that the intensity parameters $\lambda^{(\mathrm{S})}$ are smaller, i.e., we generate the data in such a way that that there are fewer extreme observations at the boundaries. Graphical convergence diagnostics or more sophisticated ones such as the Gelman–Rubin diagnostic show that convergence is not perfectly reached with 100000 iterations. Results corresponding to the more complex threshold $c_{\mathrm{S}}(t) = 1.8 + 0.1t + 0.05t^2$ and to the cases $n = 10000$ and $n = 20000$ are given in Supporting Information, as well as a surface representation of the conditional density evaluated over all possible values of the regressor.

6. Extreme forest temperatures in Switzerland

6.1. Data

The data were gathered from the Long-term Forest Ecosystem Research (LWF) database and consist of daily meteorological measurements from 14 forest weather stations, whose locations are represented in Figure 1; further information about this data may be found at:

`http://www.wsl.ch`

Measurements have been made, every ten minutes since 1997, for each site at two nearby weather stations, one under forest cover and one in the open. We keep daily average temperatures resulting in 28 series of 3414 observations. As we are primarily interested in evaluating the impact of extreme temperatures on tree species, we only keep series of measurements made under the forest canopy. We are mostly interested in the dynamics of extreme temperatures in the time domain, and hence we use the covariate time.

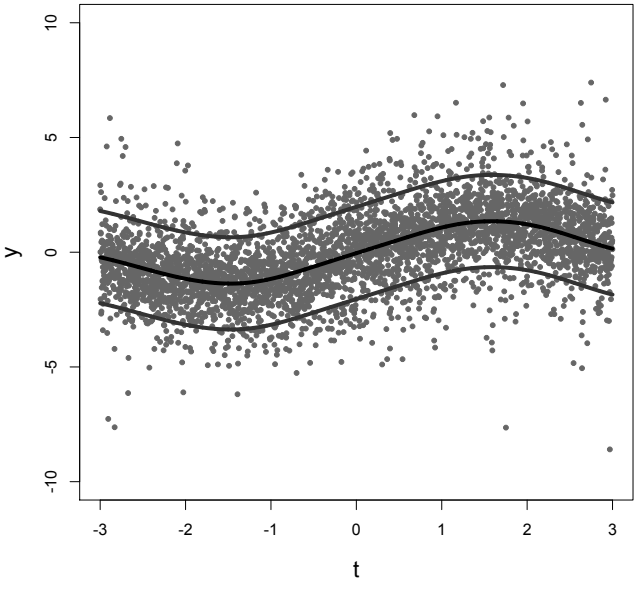


Figure 4: The solid black line represents the smooth estimate for the linear predictor of the bulk for the data in Figure 3 (a) using our Bayesian P-spline mixture model with 15 knots and second order difference penalties; the dark solid lines correspond to the estimated thresholds.

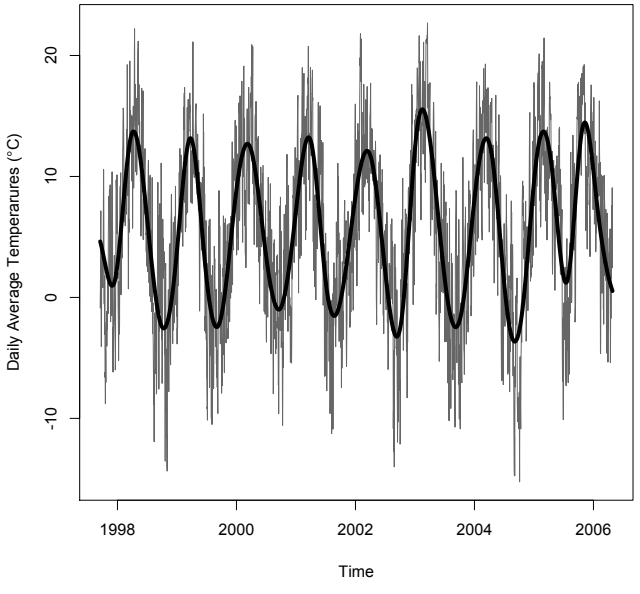


Figure 5: Smooth estimates of the bulk's mean for the data in Figure 1 (a). The gray line represents the estimate with Bayesian P-spline mixture model with 35 knots and second order difference penalties.

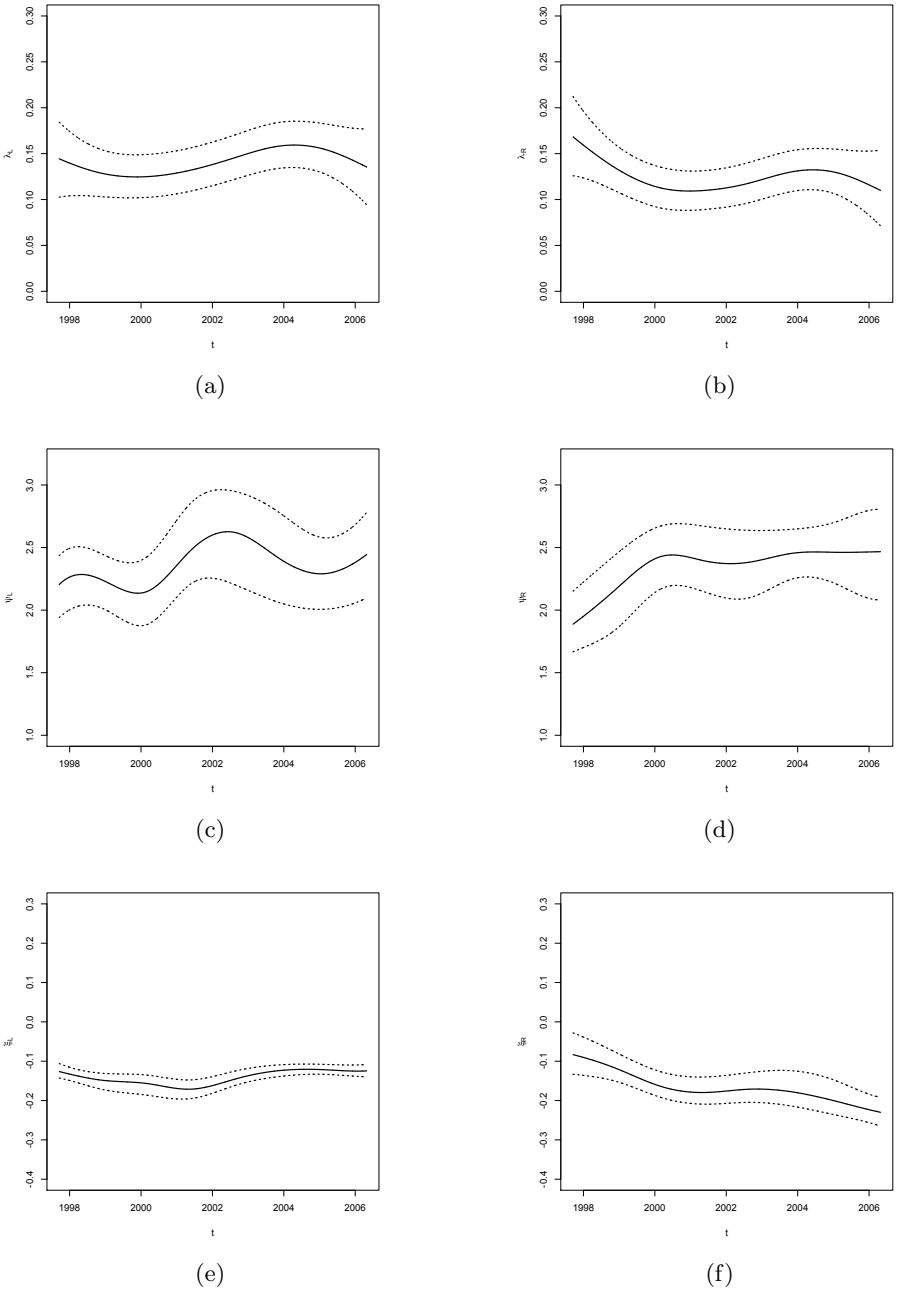


Figure 6: Smooth estimates of the tails parameters for the data in Figure 3 (a). The solid gray line represents the simulated function and the solid black line represents its estimate using our Bayesian P-spline mixture model with 15 knots and second order difference penalties; the dotted lines represent the 95% credible bands for the estimated functions.

6.2. Discussion on results

To save space, we only present here the results are presented only for Beatenberg's forest weather station; the results for the remainder forests are presented in Supporting Information. We use cubic P-splines with second order penalties, 35 knots for the bulk and 12 knots for the tail parameters. We assume that the thresholds form an envelop around the linear predictor of the bulk according to (7), and we fit a Gaussian distribution to the bulk. The smooth estimate corresponding of the linear predictor of the bulk is presented in Figure 5, and those corresponding to the tails parameters are presented in Figure 6. As it can be readily observed our smooth estimator is able to capture some interesting features of the dynamics of temperatures over the period under analysis, such as the exceptionally hot summer of 2003, which has been widely documented and studied (cf. Renaud and Rebetez, 2009). The right tail parameters seem however to be unresponsive to this event which could seem surprising on a first analysis, but this seems however to be due to the fact that since this was a prolonged period of abnormally hot weather, average temperatures have increased substantially, while the temperature levels attained during days even hotter than these were relatively low. Smooth functions for the tail parameters seem to be more responsive in years where some severe unexpected changes in temperature occurred.

7. Discussion

This paper describes a flexible approach to modeling nonstationary extremes. The proposed mixture model has the advantage of modeling simultaneously the bulk of the distribution and the parameters of the point process model for the two tails. Bayesian P-spline generalized additive models capture nonlinear smooth effects of different covariates on these parameters. The threshold is included in the model as a parameter to be estimated, thus overcoming the problems of threshold selection and uncertainties in the inference process.

Though this approach is general and permits to capture many kinds of behaviour compatible with a wealth of data generating processes, it has some defaults. First, the number of parameters to be estimated is large, convergence of the MCMC algorithm can be quite long, and it would be valuable to develop more efficient sampling algorithms. Moreover, a certain number of tail observations is necessary to obtain good estimates. The specification that the thresholds form an envelop around the linear predictor of the bulk is relatively strong but can be relaxed by considering nonlinear adjustments. Another point is that the density defined by the mixture model (5) is not necessarily continuous at the threshold, and we usually observe a jump at least at one of these two points. The use of generalized additive models for the bulk can permit to obtain continuity at one of the thresholds for a good choice of the bulk's distribution, which can be an advantage for scenarios where we are only concerned with modeling one of the tails.

Acknowledgements

We are grateful to Carl Scarott, Vanda Inácio, and Tim Hanson for helpful comments and suggestions. M. de C. acknowledges financial support from *Centro de Matemática e Aplicações, Universidade Nova de Lisboa.*

Appendix: Blockwise Metropolis–Hastings Algorithm

The sampling algorithm for simulation from the posterior of Θ is done through a blockwise Metropolis–Hastings algorithm. The proposal variances are specified to ensure an appropriate acceptance rate result for the marginal posteriors. Let $\Theta^{(0)}$ denote a starting value. For $t = 1, 2, \ldots$, suppose that

at iteration $(t-1)$, the chain is positioned at $\Theta^{(t-1)}$, and let Θ_{sf} denote parameters updated so far. The algorithm then iterates as follows:

Threshold parameters: $(c_{\text{L}}, c_{\text{R}})$.

1. Generate $c_{\text{S}}^{\star} \sim \text{LN}(\log c_{\text{S}}^{(t-1)}, \chi_{c_{\text{S}}})$; define corresponding $\mathcal{B}^{\star}, \mathcal{S}^{\star}$, and $p^{\star}(c_{\text{S}}^{\star} \mid \cdot)$;
2. Compute
$$P_{c_{\text{S}}} = \min\left\{ \frac{p^{\star}(c_{\text{S}}^{\star} \mid \Theta_{\text{sf}})}{p(c_{\text{S}}^{(t-1)} \mid \Theta_{\text{sf}})} \cdot \frac{\text{LN}(c_{\text{S}}^{(t-1)} \mid \log c_{\text{S}}^{\star}, \chi_{c_{\text{S}}})}{\text{LN}(c_{\text{S}}^{\star} \mid \log c_{\text{S}}^{(t-1)}, \chi_{c_{\text{S}}})}; 1 \right\};$$
3. With probability $P_{c_{\text{S}}}$, accept $c_{\text{S}}^{\star}$ and set $c_{\text{S}}^{(t)} = c_{\text{S}}^{\star}$; otherwise reject and set $c_{\text{S}}^{(t)} = c_{\text{S}}^{(t-1)}$. Update $\Theta_{\text{sf}}, \mathcal{L}, \mathcal{B}$, and $\mathcal{R}$.

Adjacent regression coefficients: $(\theta_l^{(j)})^{(t)}$, $l = 1, \ldots, 7$, $j = 1, \ldots, p$.

1. Generate $\theta_l^{(j)\star} \sim \text{N}((\theta_l^{(j)})^{(t-1)}, \chi_l^{(j)})$; if $l = 7$, define corresponding $\mathcal{L}^{\star}, \mathcal{B}^{\star}, \mathcal{R}^{\star}$, and $p^{\star}(\theta_7^{(j)\star} \mid \Theta_{\text{sf}})$;
2. Compute
$$P_l^{(j)} = \min\left\{ \frac{p^{\star}(\theta_l^{(j)\star} \mid \Theta_{\text{sf}})}{p((\theta_l^{(j)})^{(t-1)} \mid \Theta_{\text{sf}})}; 1 \right\};$$
3. With probability $P_l^{(j)}$, accept $\theta_l^{(j)\star}$ and set $(\theta_l^{(j)})^{(t)} = \theta_l^{(j)\star}$; otherwise reject and set $(\theta_l^{(j)})^{(t)} = (\theta_l^{(j)})^{(t-1)}$. Update $\Theta_{\text{sf}}, \mathcal{L}, \mathcal{B}$, and $\mathcal{R}$.

Fixed effect parameters: $\Lambda_l^{(t)}$, $l = 1, \ldots, 7$.

1. Generate $\Lambda_l^{\star} \sim \text{N}(\Lambda_l^{(t-1)}, \chi_l)$; if $l = 7$, define corresponding $\mathcal{L}^{\star}, \mathcal{B}^{\star}, \mathcal{R}^{\star}$, and $p^{\star}(\Lambda_7^{\star} \mid \Theta_{\text{sf}})$;
2. Compute
$$P_l = \min\left\{ \frac{p^{\star}(\Lambda_l^{\star} \mid \Theta_{\text{sf}})}{p(\Lambda_l^{(t-1)} \mid \Theta_{\text{sf}})}; 1 \right\};$$
3. With probability P_l, accept $\Lambda_l^{\star}$ and set $\Lambda_l^{(t)} = \Lambda_l^{\star}$; otherwise reject and set $\Lambda_l^{(t)} = \Lambda_l^{(t-1)}$. Update $\Theta_{\text{sf}}, \mathcal{L}, \mathcal{B}$, and $\mathcal{R}$.

Variance parameters: $(\tau_{jl}^2)^{(t)}$, $l = 1, \ldots, 7$, $j = 1, \ldots, p$.

1. Generate $(\tau_{jl}^2)^{\star} \sim \text{LN}(\log(\tau_{jl}^2)^{(t-1)}), \chi_{\tau})$;
2. Compute
$$P_{\tau} = \min\left\{ \frac{p((\tau_{jl}^2)^{\star} \mid \Theta_{\text{sf}})}{p((\tau_{jl})^{(t-1)} \mid \Theta_{\text{sf}})} \cdot \frac{\text{LN}((\tau_{jl})^{(t-1)} \mid \log(\tau_{jl}^2)^{\star}, \chi_{\tau})}{\text{LN}((\tau_{jl}^2)^{\star} \mid \log(\tau_{jl})^{(t-1)}, \chi_{\tau})}; 1 \right\};$$
3. With probability P_{τ}, accept $(\tau_{jl}^2)^{\star}$ and set $\tau_{jl}^{(t)} = (\tau_{jl}^2)^{\star}$; otherwise reject and set $\tau_{jl}^{(t)} = \tau_{jl}^{(t-1)}$. Update Θ_{sf}.

References

Behrens, C. N. and Lopes, H. F. and Gamerman, D. (2004) Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, **4**, 227–243.

Brezger, A. and Steiner, W.J. (2008) Monotonic regression based on bayesian P-splines. *Journal of Business and Economic Statistics*, **4**, 90–104.

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

Chavez-Demoulin, V. and Davison, A. C. (2005) Generalized additive modelling of sample extremes. *J. R. Statist. Soc. C*, **54**, 207–222.

Davison, A. C. and Ramesh, N. I. (2000) Local likelihood smoothing of sample extremes. *J. R. Statist. Soc. B*, **62**,191–208.

Davison, A. C. and Smith, R. (1990) Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc. B*, **52**, 393–442.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–102.

Ferrez, J. and Davison, A. C. and Rebetez, M. (2011) Extreme temperature analysis under forest cover compared to an open field. *Agricultural and Forest Meteorology*, **151**, 992–1001.

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

Hall, P. and Tajvidi, N. (2000) Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, **15**, 153–167.

Hastie, T. J. and Tibshirani, R. J. (1990) Generalized additive models. *London: Chapman & Hall.*

Lang, S. and Brezger, A. (2004) Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Laurini, F. and Pauli, F. (2009) Smoothing sample extremes: the mixed model approach. *Computational Statistics & Data Analysis*, **53**, 3842–3854.

Leadbetter, M. R. and Lindgren, G. and Rootzen, H. (1983) Extremes and related properties of random sequences and processes.. *Z. Wahrsch. Ver. Geb*, **65**, 291–306.

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

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

Northrop, P. J. and Jonathan, P. (2011) Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics* (DOI:10.1002/env.1106).

Padoan, S. A. and M.P. Wand (2008) Mixed model-based additive models for sample extremes. *Statist. Probab. Lett.*, **78**, 2850–2858.

Ramesh, N. I and Davison, A. C. (2002) Local models for exploratory analysis of hydrological extremes. *Journal of Hydrology*, **256**, 106–119.

Renaud V., Rebetez M. (2009) Comparison between open-site and below-canopy climatic conditions in Switzerland during the exceptionally hot summer of 2003. *Agricultural and Forest Meteorology*, **149**, 873–880.

Tancredi, A. and Anderson, C. and O'Hagan, A. (2006) Accounting for threshold uncertainty in extreme value estimation. *Extremes*, **9**, 87–106.