Variables selection in a P-spline regression model for flood frequency analysis

A. Gardini[1,*]

[1] *Department of Statistical Sciences; University of Bologna, Italy; aldo.gardini@unibo.it*
[*] *Corresponding author*

Abstract. *The prediction of occurrence probability of extreme environmental phenomena is one of the most challenging tasks for a statistician. In this paper, we tackle the problem of estimating return levels of river discharge, relevant in flood frequency analysis, by relying on extreme value theory. The Generalized Extreme Value (GEV) distribution is assumed to model the sequences of river discharge annual maxima observed at multiple gauging stations belonging to the same river basin. A unique Bayesian model is specified, with the aim of investigating how covariates characterizing stations sub-catchments can be useful to estimate return levels, also at ungauged locations. As non-linear relationships may occur among GEV parameters and covariates, Bayesian P-splines are included in the latent models and the problem of variable selection is faced. In particular, a grouped HorseShoe prior is specified on splines coefficients, to encourage the shrinkage of non-relevant components to zero. Results from a real data application on the Upper Danube basin show the potential of the proposed method in reducing the estimates uncertainty.*

Keywords. *Bayesian inference; Generalized extreme value distribution; HorseShoe prior; Return levels; Stan.*

---

1 Introduction

The prediction of return period of extreme flood events is crucial for several aspects related to land management [11, 7]. To this aim, measures of river discharge recorded at different time periods from gauging stations can be useful for making inference on flood events. Extreme value theory [3] is the natural branch of statistics in which to perform flood frequency analysis. Among the different approaches, the analysis of a block-maxima diver discharge sequence (year maxima in the application) through the Generalized extreme value (GEV) distribution is faced. More formally, maxima discharge values related to T distinct years are denoted by y_t, $t = 1,\dots,T$, and, for each observation, it is assumed that $y_t|\mu,\sigma,\xi \overset{ind}{\sim} GEV(\mu,\sigma,\xi)$, $\forall t$. Such a distribution is ruled by three parameters: $\mu \in \mathbb{R}$ controls the location, $\sigma \in \mathbb{R}^+$ controls the scale and $\xi \in \mathbb{R}$ controls the shape, affecting the behaviour of the distribution tails. In the statistical analysis of extremes, the most typical output is the estimation of return levels associated with a return period R. They are defined as the quantile Q_p that has a probability equal to $p = 1/R$ of being exceeded at each year.

An important outcome of flood frequency analysis is to provide reliable estimates for return levels also at ungauged locations, taking advantage on the presence of covariates and/or delineating homoge-

neous regions. These approaches falls within the regional flood frequency analysis framework [6] and the most classical one is the flood index method [4]. In this context, a Bayesian GEV regression model on data from multiple stations available in the studied basin can be built, guaranteeing uncertainty propagation [11]. Furthermore, such an approach allows to taking advantage on the use of covariates, that are crucial to obtain reliable estimates of the GEV parameters for ungauged location. It might be observed that the relationships among GEV parameters (or their functions) and available covariates are not linear, requiring the specification of flexible regression models by considering, for example, spline terms. In this paper, Bayesian P-splines are included in the model and the specification of a group HorseShoe (HS) prior is discussed, in order to let the model automatically perform the variable selection step. Indeed, such a task is not trivial in this modelling framework, dealing with three distinct parameters not always easy to interpret.

The rest of the paper is organized as follows: the dataset on Danube river basin that motivated the analysis is described in Section 2 and the developed statistical model in Section 3. Section 4 reports some results from the real data application and, lastly, Section 5 offers some concluding remarks.

2 Data on Danube basin

The application that motivated the developed statistical model targets the problem of estimating return levels for the discharge of rivers belonging to the Danube upper basin (i.e. the part located both in Germany and Austria). To propose a general procedure, the analysis considers only data that are publicly available. In this way, it is possible to easily replicate the steps of the statistical method also in other river basins. The time series with daily river discharge data are retrieved from the GRDC portal [10]. The sub-catchments of each station, together with useful covariates (area, mean slope, mean elevation, and aspect direction) can be computed starting from the EU-DEM [5]. Additional covariates characterizing the rainfall in each sub-catchment can be computed from the bioclimatic variables rasters [8]. Lastly, the proportion of area covered by buildings is computed from land cover rasters [2]. At the end of the procedure, 63 gauging stations in the Upper Danube basin are included in the analysis, considering yearly maxima from 1985 to 2017.

3 The proposed statistical model

Let us consider that a collection of N maxima y_{st} from $s = 1, \dots, S$ gauging stations are available for blocking times $t = 1, \dots, T$. It is assumed that, conditionally on site specific covariates, the maxima are independently distributed as:

$$y_{st} | \mu_s, \sigma_s, \xi_s \overset{ind}{\sim} GEV(\mu_s, \sigma_s, \xi_s). \tag{1}$$

The assumption of conditional independence represents a simplification, but it is quite standard in the extreme value literature if marginal return levels need to be estimated [11, 7]. Alternatively, max-stable spatial processes can be considered, even if the complexity of the modelling framework sensibly increases [1]. The multivariate link function proposed in [7] is considered to obtain transformed parameters which is convenient to specify regression models for:

$$g(\mu_s, \sigma_s, \xi_s) = (\psi_s = \log(\mu_s), \tau_s = \log(\sigma_s/\mu_s), \phi_s = h(\xi_s))^\top .$$

The logarithmic transformation applied to the location parameter μ_s restricts its domain to $\mathbb{R}^+$, according to the exploratory results, such assumption is fulfilled in the discussed application; furthermore, the dependence between scale and location parameters is mitigated by modeling $\log(\sigma_s/\mu_s)$ instead of the most common choice $\log(\sigma_s)$. Lastly, the function applied to the shape parameter is

$$h(\xi_s) = a_\phi + b_\phi \log\{-\log[1-(\xi+0.5)^{c_\phi}]\}, \tag{2}$$

where $(a_\phi, b_\phi, c_\phi) = (0.062376, 0.39563, 0.8)$. Johannesson and colleagues [7] proposed it in order to restrict the domain of ξ_s to the interval $(-0.5, 0.5)$, with the aim of keeping a linear relationship with ξ_s around zero. The domain restriction for the shape parameter is motivated by the desiderata of guaranteeing the variance of the GEV distribution to be finite and its upper bound to be greater than $\mu + 2\sigma$. It is highlighted that the developments discussed later still holds, even if a different link function is adopted. After defining the vectors of station-specific parameters $\psi = (\psi_1, \ldots, \psi_S)^\top$, $\tau = (\tau_1, \ldots, \tau_S)^\top$ and $\phi = (\phi_1, \ldots, \phi_S)^\top$, the following system of latent regression models is set:

$$\begin{aligned}
\psi &= \mathbf{1}_S \beta_{0\psi} + \sum_{m=1}^{M} \mathbf{B}_m \gamma_{\psi,m} + \mathbf{u}_\psi; \\
\tau &= \mathbf{1}_S \beta_{0\tau} + \sum_{m=1}^{M} \mathbf{B}_m \gamma_{\tau,m} + \mathbf{u}_\tau; \\
\phi &= \mathbf{1}_S \beta_{0\phi} + \sum_{m=1}^{M} \mathbf{B}_m \gamma_{\phi,m} + \mathbf{u}_\phi.
\end{aligned} \tag{3}$$

Each linear predictor, related to the generic parameters vector $\theta \in \{\psi, \tau, \phi\}$, is constituted by an overall intercept $\beta_{0\theta}$, the sum of M flexible regression terms defined as the product of a matrix of cubic B-spline basis functions evaluated at K knots, $\mathbf{B}_m \in \mathbb{R}^{S\times K}$, multiplied by a vector of associated coefficients $\gamma_{\theta,m} \in \mathbb{R}^K$. A vector of station-specific unstructured random effects $\mathbf{u}_\theta \in \mathbb{R}^S$ complements the equation, to account for possible residual variation. To keep the notation simple, model equations in (3) contain the same regression terms, as will be the case in the application, but this assumption can be easily relaxed.

To complete the model specification, prior distribution on the parameters must be set. Firstly, a weakly informative Gaussian prior is assumed for the intercepts, whereas for the unstructured random effects an uncorrelated multivariate Gaussian prior with scale parameter $\sigma_{u\theta}$ is set:

$$\mathbf{u}_\theta | \kappa_\theta \sim \mathcal{N}_S(\mathbf{0}, \kappa_\theta^2 \mathbf{I}_S), \quad \kappa_\theta \sim \mathcal{N}^+(0, 2^2), \quad \theta \in \{\psi, \tau, \phi\};$$

where $\mathcal{N}^+(\cdot,\cdot)$ indicates an half-Normal distribution. The crucial step is establishing priors for the spline coefficients vectors. To encourage smoothness, a P-spline approach is adopted, selecting the priors

$$\gamma_{\theta,m} | \omega_{\theta,m} \sim \mathcal{N}_K(\mathbf{0}, \omega_{\theta,m}^2 \mathbf{K}_\gamma^-), \quad \omega_{\theta,m} \sim \mathcal{N}^+(0, 2^2), \quad m = 1, \ldots, M;\ \theta \in \{\psi, \tau, \phi\};$$

where $\mathbf{K}_\gamma$ is a precision matrix describing a second-order random walk prior and has rank $K-2$.

Exploiting the spectral decomposition of $\mathbf{B}_m \mathbf{K}_\gamma^- \mathbf{B}_m^\top$, i.e. the covariance matrix of $\mathbf{B}_m \gamma_{\theta,m}$, it is possible to reparametrize the model. In this way, the structured and improper prior on the spline coefficients are traced back to proper iid Gaussian priors [9]. Without reporting the technical details, the equation models can be re-expressed as:

$$\theta = \mathbf{1}_S \beta_{0\theta} + [\mathbf{x}_1 \cdots \mathbf{x}_M] \beta_\theta + \sum_{m=1}^{M} \tilde{\mathbf{B}}_m \tilde{\gamma}_{\theta,m} + \mathbf{u}_\theta, \tag{4}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are the centred numerical covariates and $\beta_\theta \in \mathbb{R}^M$ the vector of related linear coefficients, $\tilde{\mathbf{B}}_m \in \mathbb{R}^{S\times(K-2)}$ is a matrix of orthonormal basis functions and $\tilde{\gamma}_{\theta,m} \in \mathbb{R}^{K-2}$ is the associated vector of coefficients, for which now the prior is $\tilde{\gamma}_{\theta,m} | \omega_{\theta,m} \sim \mathcal{N}_{K-2}(\mathbf{0}, \omega_{\theta,m}^2 \mathbf{I}_{K-2})$. Under this parametrization, if a weakly informative prior is assumed on β_θ, then a standard P-splines model can be implemented.

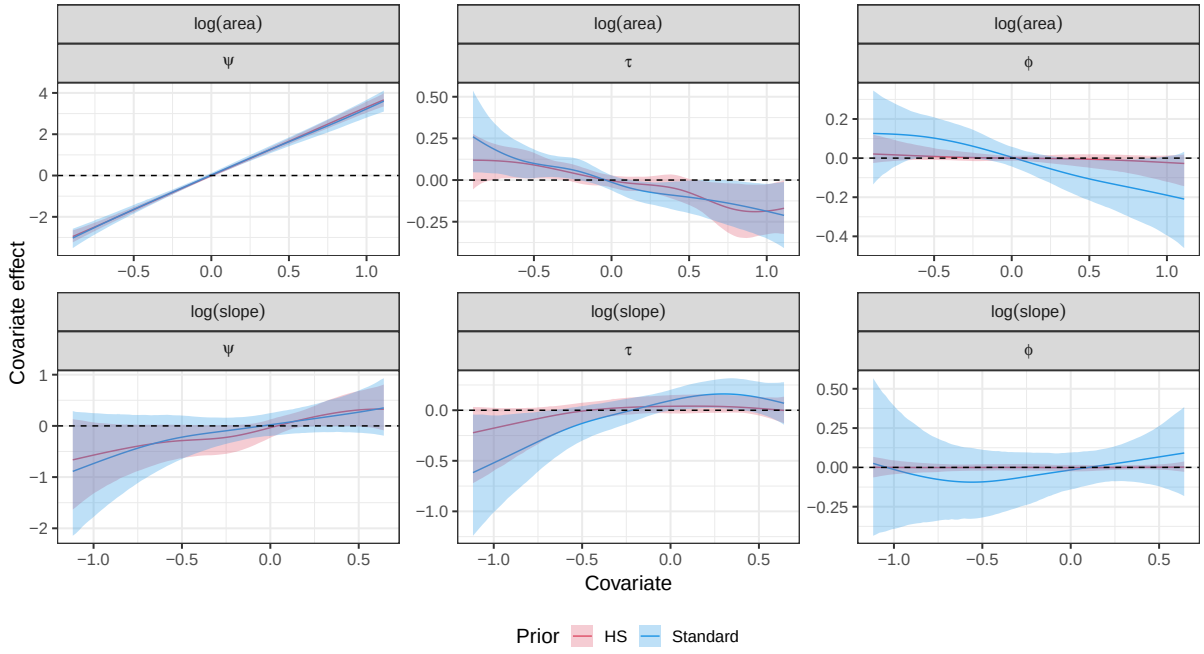

Figure 1: Estimated effects of two covariates in the three models on latent parameters. The solid line represents the posterior mean and the shaded area the 90% credible interval.

3.1 Variable selection: the Grouped HorseShoe prior

When several covariates are available and their relationships with the modelled latent parameters are unknown, it can be useful to set a prior distribution that is able to shrink the non-relevant regressors to zero. Scheipl and colleagues [9] proposed to use spike-and-slab priors for functional selection. The behaviour of such priors is also mimicked by the HS priors, that have been proposed in a hierarchical version to deal with shrinkage of grouped regression terms [12]. Since all the coefficients related to a spline term can be considered to form a group, a grouped HS prior appear suitable to be applied in this framework, slightly reparametrizing the model in (4) to

$$\boldsymbol{\theta} = \mathbf{1}_S \beta_{0\theta} + \sum_{m=1}^{M} \mathbf{Z}_m \boldsymbol{\alpha}_{\theta,m} + \mathbf{u}_\theta, \tag{5}$$

where $\mathbf{Z}_m = [\mathbf{x}_m \;\; \tilde{\mathbf{B}}_m] \in \mathbb{R}^{S\times(K-1)}$. To implement the grouped HS prior, the following hierarchy is necessary:

$$\begin{aligned}
&\boldsymbol{\alpha}_{\theta,m} | \delta_{1m}, \ldots, \delta_{(K-1)m}, \lambda_m, \eta \sim \mathcal{N}_{K-1}(\mathbf{0}, \eta^2 \lambda_m^2 \text{diag}[\delta_{1m} \cdots \delta_{(K-1)m}]) \; m = 1, \ldots M; \\
&\delta_{km} \sim \mathcal{C}^+(0,1), \;\; k = 1, \ldots, K-1; \;\; m = 1, \ldots M; \\
&\lambda_m \sim \mathcal{C}^+(0,1), \; m = 1, \ldots M; \\
&\eta \sim \mathcal{C}^+(0,1);
\end{aligned} \tag{6}$$

where $\mathcal{C}^+(\cdot,\cdot)$ denotes an half-Cauchy distribution. Since all the latent variables involved in the prior are continuous, in contrast with spike-and-slab priors, the popular `Stan` software was exploited to draw samples from the posterior distribution of target parameters.

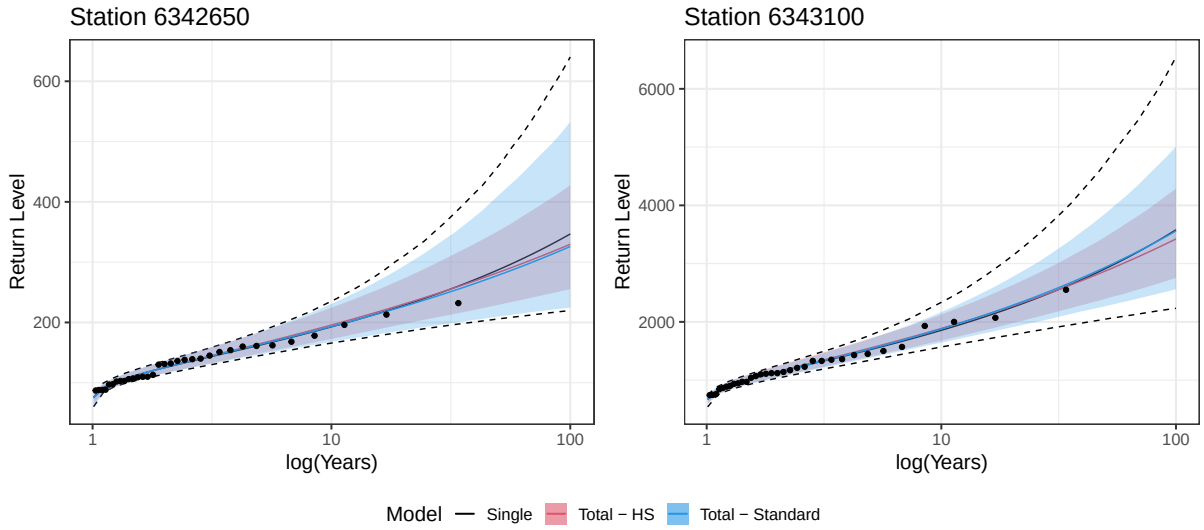

Figure 2: Return levels estimated for two stations in the studied basin. Solid lines represent the posterior means, the shaded areas and the dashed lines report the 90% credible intervals.

4 Some results

The statistical model described in the previous section is fitted on the $N = 2046$ yearly maxima recorded in the 62 studied stations, including $M = 8$ covariates: coordinates of the station, the sub-catchment area, elevation, slope, average aspect, proportion of build area and rainfall. The model is considered both under standard priors for P-splines coefficients (labelled as *Standard*) and under the proposed grouped HS prior.

The effectiveness of grouped HS prior in shrinking functional relationships is depicted in Figure 1. Focusing on the effect of sub-catchment area covariate on ψ, it can be noted that the HS prior shrinks to 0 splines coefficients, reducing it to a linear one, whereas standard priors still shows a slightly non-linear behaviour. In general, credible intervals are narrower when the HS prior is adopted, and this is particularly evident in the model on ϕ, i.e. the transformation of the GEV shape parameter. It is widely known that identifying this parameter is tricky [7], and if a strong signal is not detected, HS priors are able to penalise related coefficients.

A direct consequence of the posterior variability reduction for shape parameters can be noted in Figure 2, where the fitted return levels are reported with respect to the corresponding return time. As a benchmark, results of the GEV model fitted on the single station areas reported too. The red shaded are showing that 90% credible intervals under the overall basin model with grouped HS prior are narrower that the others. The differences are more evident when the return time increase, probably due to less posterior variances of the GEV shape parameters.

5 Concluding remarks

The paper focuses on the problem of integrating covariates in extreme value models for flood frequency analysis, due to their relevance in estimating return levels for ungauged locations. Non-linear relation-

ships may occur among GEV parameters and covariates, therefore flexible P-splines terms are included in the model. The proposed model is completed by the adoption of grouped HS priors, to address the variable selection step. All the models are implemented in `Stan`, to foster their accessibility also to practitioners. Concerning the weakly informative priors, their hyper-parameters are calibrated exploiting results from preliminary analyses: in this way, they account for the different scales characterizing the GEV parameters. Of course, further investigations on the capability of the proposed model in predicting return levels for out-of-sample locations should be carried out.

Acknowledgments. The work of Aldo Gardini was partially supported by MUR on funds FSE REACT EU - PON R&I 2014-2020 and PNR (D.M. 737/2021) for the RTDA_GREEN project (title: "Modelli statistici per lo studio della convergenza spaziale verso la transizione verde", J41B21012140007).

References

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

[2] Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N. E., Herold, M., and Fritz, S. (2020). Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe 2020. https://land.copernicus.eu/global/products/lc

[3] Coles, S. (2001). An introduction to statistical modeling of extreme values. Springer-Verlag.

[4] Dalrymple, T. (1960). Flood-frequency analyses, manual of hydrology: Part 3 (No. 1543-A). USGPO.

[5] European Environment Agency - Copernicus Programme. European Digital Elevation Model (EU-DEM), version 1.1. https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1

[6] Hosking, J., and Wallis, J. (1997). Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge: Cambridge University Press.

[7] Johannesson, A. V., Siegert, S., Huser, R., Bakka, H., and Hrafnkelsson, B. (2022). Approximate bayesian inference for analysis of spatiotemporal flood frequency data. *The Annals of Applied Statistics*, **16**(2), 905?935.

[8] O'Donnel, M. S., Ignizio, D. A. (2012). Bioclimatic predictors for supporting ecological applications in the conterminous United States. *US geological survey data series*, **691**, 4-9.

[9] Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, **107**(500), 1518?1532.

[10] The Global Runoff Data Centre, 56068 Koblenz, Germany. (1988). The world-wide repository of river discharge data and associated metadata. https://portal.grdc.bafg.de/

[11] Thorarinsdottir, T. L., Hellton, K. H., Steinbakk, G. H., Schlichting, L., and Engeland, K. (2018). Bayesian regional flood frequency analysis for large catchments. *Water Resources Research*, **54**(9), 6929?6947.

[12] Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). Bayesian grouped horseshoe regression with application to additive models. In *AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29*, pages 229?240. Springer.