Reparameterization of extreme value framework for improved Bayesian workflow

Théo Moins, Julyan Arbel, Stéphane Girard, Anne Dutfoy

Reparameterization of extreme value framework for improved Bayesian workflow

Théo Moins * Julyan Arbel * Stéphane Girard * Anne Dutfoy †

October 7, 2022

Abstract

Combining extreme value theory with Bayesian methods offers several advantages, such as a quantification of uncertainty on parameter estimation or the ability to study irregular models that cannot be handled by frequentist statistics. However, it comes with many options that are left to the user concerning model building, computational algorithms, and even inference itself. Among them, the parameterization of the model induces a geometry that can alter the efficiency of computational algorithms, in addition to making calculations involved. We focus on the Poisson process characterization of extremes and outline two key benefits of an orthogonal parameterization addressing both issues. First, several diagnostics show that Markov chain Monte Carlo convergence is improved compared with the original parameterization. Second, orthogonalization also helps deriving Jeffreys and penalized complexity priors, and establishing posterior propriety. The analysis is supported by simulations, and our framework is then applied to extreme level estimation on river flow data.

1 Introduction

Studying the long-term behavior of environmental variables is necessary to understand the risks of hazardous meteorological events such as floods, storms, or droughts. To this end, models from extreme value theory allow to extrapolate data in the tails of the distribution, in order to estimate extreme quantiles that may not have been observed (see Coles, 2001, for an introduction). In particular, key quantities to estimate are return levels ℓ_T associated with a given period of T years. They correspond to the level that is exceeded in average once every T years. Assessing the resistance of facilities to natural disasters such as dams to floods that occur in average once every 100 years or 1 000 years is critical for companies like Électricité de France (EDF). Moreover, characterizing the uncertainty on the estimation of this return level is also of interest, which encourages the choice of the Bayesian paradigm over the frequentist one. However, doing Bayesian inference requires multiple steps that must be managed by the user, from the choice of the model to the evaluation and validation of computations. This has been recently formalized by Gelman et al. (2020) in the form of a Bayesian workflow. After introducing models stemming from extreme value theory in Section 1.1, we briefly review in Section 1.2 one particular step of the workflow which is reparameterization, and more specifically the choice of an orthogonal parameterization.

*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
†EDF R&D dept. Périclès, 91120 Palaiseau, France.

1.1 Extreme-value models

Three different frameworks exist to model extreme events, leading to different likelihoods: one by block maxima, one by peak-over-threshold, and one that unifies both by a Poisson process characterization.

Block maxima model Let X be a random variable with cumulative distribution function (cdf) F, and M_n the maximum of n i.i.d random variables with cdf F, whose cdf is consequently F^n. We consider the case where F belongs to a maximum domain of attraction, which means that there exist two sequences $a_n > 0$, b_n and a cdf G such that $F^n(a_n x + b_n) \to G(x)$ as $n \to \infty$. The extreme value theorem (see for instance Haan and Ferreira, 2006) states that G is necessarily a generalized extreme-value (GEV) distribution, with cdf:

$$G(x) = \begin{cases} \exp\left(-\{1+\xi x\}_+^{-1/\xi}\right) & \text{if } \xi \neq 0\,, \\ \exp(-\exp(-x)) & \text{if } \xi = 0, \end{cases} \tag{1}$$

with $\xi \in \mathbb{R}$ and where $\{x\}_+ = \max\{0, x\}$. Consequently, for a finite value of n, one can consider the approximation $\mathbb{P}(M_n \leq x) \approx G((x-b_n)/a_n) =: G(x \mid b_n, a_n, \xi)$, and focus on the estimation of the three resulting parameters of the GEV distribution. To obtain a sample of maxima, the common approach by block maxima consists in dividing the dataset into m blocks of size n/m and extract the maximum from each of them.

Peak-over-threshold model Alternatively, one can consider observations that exceed a high threshold u. The second extreme value theorem, also known as Pickands theorem (Pickands, 1975) states that, if F belongs to the maximum domain of attraction of G with $\mathbb{P}(M_n \leq x) \approx G(x \mid \mu, \sigma, \xi)$, then the distribution of the exceedances $X - u \mid X > u$ is asymptotically, as u converges to the endpoint of F, a generalized Pareto distribution (GPD), with cdf:

$$H(y \mid \tilde{\sigma}, \xi) = \begin{cases} 1 - \left\{1 + \xi \frac{y}{\tilde{\sigma}}\right\}_+^{-1/\xi} & \text{if } \xi \neq 0\,, \\ 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right) & \text{if } \xi = 0, \end{cases} \tag{2}$$

where the shape parameter ξ is the same as in (1) and the relation $\tilde{\sigma} = \sigma + \xi(u - \mu)$ links the GPD and GEV scales. Here, to obtain a sample of k excesses, the peak-over-threshold method consists in choosing u as the $(n-k)$th order statistic and consider only the k largest values of the dataset. This method thus requires the estimation of the quantile of order $1 - k/n$, which can be seen as the third parameter to estimate, in addition to $\tilde{\sigma}$ and ξ.

Poisson process characterization of extremes Finally, these two approaches can be generalized by a third one, using a non-homogeneous Poisson process. We present here an intuitive way for obtaining this model similarly to (Coles, 2001, Chapter 7), and refer to Leadbetter et al. (1983) for more details on point process theory and technical details associated with this construction. Here, we start by observing that, for large n, $F^n(x) \approx G(x \mid \mu, \sigma, \xi)$, for x in the support of G denoted by $\text{supp}(G(\cdot \mid \mu, \sigma, \xi)) = \left\{x \in \mathbb{R} \quad \text{s.t. } 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$. Hence, considering a large threshold $u \in \text{supp}(G(\cdot \mid \mu, \sigma, \xi))$, a Taylor expansion yields $n \log F(u) \simeq -n(1 - F(u)) \simeq \log G(u \mid \mu, \sigma, \xi)$, or, equivalently,

$$\mathbb{P}\left(X > u\right) \simeq -\frac{1}{n} \log G(u \mid \mu, \sigma, \xi). \tag{3}$$

Equation (3) can be seen as the probability of the random variable X to belong to $I_u := [u, +\infty)$. In the case of n i.i.d random variables, we can deduce that the associated point process N_n is such that $N_n(I_u) \sim \mathcal{B}(n, p_n)$ with p_n given in Equation (3). As $n \to +\infty$, the binomial distribution $\mathcal{B}(n, p_n)$ converges to the Poisson distribution $\mathcal{P}(\Lambda(I_u))$, with $\Lambda(I_u) = -\log G(u \mid \mu, \sigma, \xi)$. This property being valid for all I_u together with the independence property on non-overlapping sets imply that N_n converges to a non-homogeneous Poisson process, with intensity measure $\Lambda(I_u)$: $N_n \xrightarrow{d} N$, with $N(I_u) \sim \mathcal{P}(\Lambda(I_u))$. This model generalizes the block maxima one since

$$\mathbb{P}(M_n < x) = \mathbb{P}(N_n(I_x) = 0) \to \mathbb{P}(N(I_x) = 0) = \exp(-\Lambda(I_x)) = G(x \mid \mu, \sigma, \xi), \text{ as } n \to \infty.$$

However, an estimation of the parameters (μ, σ, ξ) with this model are related to the overall maximum of the dataset M_n, and it is frequent to study maxima of m smaller blocks $M_{n/m}$, where m is typically the number of years in the observations and so $M_{n/m}$ corresponds to annual maxima. To do so, the intensity measure is multiplied by m, which modifies the parameterization and in particular (μ, σ) but not ξ: Wadsworth et al. (2010) shows that, if $(\mu_{k_1}, \sigma_{k_1}, \xi)$, resp. $(\mu_{k_2}, \sigma_{k_2}, \xi)$, are parameters for k_1, resp. k_2, block maxima, then one has:

$$\mu_{k_2} = \mu_{k_1} - \frac{\sigma_{k_1}}{\xi}\left(1 - \left(\frac{k_2}{k_1}\right)^{-\xi}\right), \quad \sigma_{k_2} = \sigma_{k_1}\left(\frac{k_2}{k_1}\right)^{-\xi}. \tag{4}$$

The threshold excess model can also be derived from the point process representation, as one can show that $\mathbb{P}(X > y + u \mid X > u) \simeq 1 - H(y \mid \tilde{\sigma}, \xi)$, with $\tilde{\sigma} = \sigma + \xi(u - \mu)$. Moreover, in contrast to the peak-over-threshold model, this one includes directly the estimation of a position parameter instead of an intermediate quantile.

In the following, we will focus mostly on this latter model, and treat the peak-over-threshold one as a special case in Section 4.1.

Bayesian inference Using the Bayesian paradigm for extremes has been shown to give several advantages over frequentist methods, such as including expert information using informative prior, not making any assumption on the value of ξ or considering the uncertainty related to the parameter estimation in the prediction of a new event thanks to the posterior predictive expression. See Coles and Powell (1996) for a general review, and Stephenson (2016) or Bousquet (2021) for more recent overviews about Bayesian extremes.

For the Poisson process characterization of extremes, Bayesian inference consists in fixing a threshold u and a scaling factor m, and then considering the n_u observations $\boldsymbol{x} = (x_1, \ldots, x_{n_u})$ that exceed u. The likelihood of these observations can be written as

$$L(\boldsymbol{x}, n_u \mid \mu, \sigma, \xi) = \exp\left(-m\left(1 + \xi\left(\tfrac{u-\mu}{\sigma}\right)\right)^{-1/\xi}\right)\sigma^{-n_u}\prod_{i=1}^{n_u}\left(1 + \xi\left(\tfrac{x_i-\mu}{\sigma}\right)\right)^{-1-1/\xi}. \tag{5}$$

A complete Bayesian model requires also the specification of a prior $p(\mu, \sigma, \xi)$, to be able to obtain the posterior $p(\mu, \sigma, \xi \mid \boldsymbol{x}, n_u)$ using Bayes' theorem, $p(\mu, \sigma, \xi \mid \boldsymbol{x}, n_u) \propto p(\mu, \sigma, \xi)L(\boldsymbol{x}, n_u \mid \mu, \sigma, \xi)$. This posterior summarizes the information on the parameters after observations, and can be used to extract point estimators, build credible intervals, or write the probability of a new observation $\tilde{x}$ given data $\boldsymbol{x}$ using the posterior predictive:

$$p(\tilde{x} \mid \boldsymbol{x}, n_u) = \int p(\tilde{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{x}, n_u)\mathrm{d}\boldsymbol{\theta}, \qquad \boldsymbol{\theta} = (\mu, \sigma, \xi). \tag{6}$$

These quantities of interest are rarely explicit, and are often derived by sampling approaches. A general overview of the Bayesian workflow is given in Gelman et al. (2020), and we focus here on the particular step of reparameterization for the likelihood $L(\boldsymbol{x}, n_u \mid \mu, \sigma, \xi)$ in the case where Markov chain Monte Carlo (MCMC) methods are used to approximate the posterior distribution.

1.2 Reparameterization

Although the choice of parameterization of a statistical model does not alter the model *per se*, it does reshape its geometry, which in turn may impact inferential aspects such as efficiency or accuracy. This is the case for Bayesian inference, and especially in MCMC strategies. For these methods, a crucial complication for convergence is parameters correlation. This notion of correlation between parameters can be associated to a notion of orthogonality, as it asymptotically leads to the independence of posterior components.

Parameterization and MCMC It has been known for several decades that parameterization is crucial for a good mixing of MCMC chains, especially with the correlation between the coordinates. Gilks et al. (1995, Chapter 6) is a great introduction for Gibbs sampling and Metropolis–Hastings algorithm. For Gibbs, highly dependent components can lead to iterations concentrated to each others, which causes a slow mixing as a lot of iterations are therefore required to explore the parameter space. An illustration is given by Hills and Smith (1992) with a bivariate Gaussian distribution, where the convergence rate is explicit and is affected by the correlation between the two components. More general computations are conducted by Roberts and Sahu (1997) in the normal case, as the dependence between coordinates can easily be modeled with correlation. However, this convergence rate is less explicit in the general case, see for example Roberts and Polson (1994). For Metropolis–Hastings, if the structure of the jumping kernel is not similar to the one of the target density (which is a typical case if there is a complex dependence between parameters), then too many candidates generated by the kernel are rejected and the same problem as for Gibbs sampling occurs. For more recent MCMC algorithms such as Hamiltonian Monte Carlo (HMC, Neal, 2011) and its variant NUTS (Hoffman and Gelman, 2014), Betancourt and Girolami (2015) gives an example of the benefit of reparameterization for hierarchical models. Another one can be found in Vehtari et al. (2021) with a Cauchy likelihood, where the issue is not due to correlations but rather to the Cauchy distribution tail-heaviness, that hinders exploration by Hamiltonian dynamics. More generally, Betancourt (2019) studies reparameterization from a geometric perspective, in order to show its equivalence with adapted versions of HMC on Riemannian manifold.

Due to the difficulty to obtain general results on reparameterization and MCMC convergence, a significant part of the research focuses on specific models, such as hierarchical models (Papaspiliopoulos et al., 2003; Browne et al., 2009), linear regression (Gilks et al., 1995), or mixed models (Gelfand et al., 1995, 1996). An overview of parameterization methods is given in Gelman (2004) in the case of data augmentation and parameters expansion. In addition to improving the MCMC convergence, Gelman (2004) shows that a good parameterization also allows to better interpreting model parameters.

Orthogonal parameterization As seen before, reducing the dependence between the coordinates is desirable for MCMC. One way to characterize dependence is with asymptotic covariance and the notion of orthogonality according to Jeffreys (1961): parameters are said to be orthogonal when the Fisher information is diagonal. With this definition, having orthogonal parameters leads

to asymptotic posterior independence when a Bernstein–von Mises theorem holds (more details on Bernstein–von Mises theorems can be found in Van der Vaart, 2000, Chapter 10). Studied by Huzurbazar (1950), the problem of finding an orthogonal parameterization is seldom feasible when there are more than three parameters, since the number of equations is then greater than the number of unknown variables. In the case of three parameters, there are as many equations as there are unknowns, but the non-linear system does not necessarily lead to a solution.

The main use of orthogonal parameterization is to make parameters of interest independent of nuisance parameters (Cox and Reid, 1987). Other definitions of orthogonality are also proposed to be more adapted to the inferential context (Tibshirani and Wasserman, 1994) or to ensure consistency of the parameter of interest (Woutersen, 2011). For Bayesian inference, Tibshirani and Wasserman (1994) compares different definitions and suggests a strong assumption of normality for the posterior. In the following, we keep the most popular definition of orthogonality due to Jeffreys (1961), as we are not interested in properties associated with the estimation of a given parameter of interest, but rather on the dependence structure between parameters. However, up to our knowledge, there is no clear evidence in the literature of a direct link between parameter orthogonality and mixing properties of the corresponding MCMC chains, such as a better convergence rate. In Section 4, we bring some empirical evidence on the interest of orthogonality in extreme value models.

1.3 Contributions and outline

In this paper, we study the benefits of reparameterization for the Poisson process characterization of extremes in a Bayesian context. In particular, we show that the orthogonal parameterization is useful for several reasons: we argue in Section 2 that it improves the performance of MCMC algorithms in terms of convergence, and we show in Section 3 that it also facilitates the derivation of priors such as Jeffreys and an informative variant on the shape parameter using penalized complexity (PC) priors (Simpson et al., 2017). These results are then illustrated by experiments in Section 4, first on simulations to compare the different parameterizations, and second on a dataset of the Garonne river flow to apply our model on real data. All the proofs as well as additional figures are provided in the Appendix, and the code corresponding to the experiments is available online.[1]

2 Reaching orthogonality for extreme Poisson process

An attempt to reparametrizing the Poisson process for extremes in order to improve MCMC convergence already exists in the literature (Sharkey and Tawn, 2017), but has several limitations that we detail here. Instead, we suggest to use the fully orthogonal parameterization of Chavez-Demoulin and Davison (2005).

Near-orthogonality with hyperparameter tuning Based on the relationship between parameters given in Equation (4), Sharkey and Tawn (2017) suggests to change the scaling factor m before using Metropolis–Hastings algorithm in order to optimize MCMC convergence. To this aim, they minimize the non-diagonal elements of the inverse Fisher information matrix corresponding to asymptotic covariances. Then, the parameters corresponding to the initial number of blocks are retrieved with Equation (4). As the calculations cannot be achieved explicitly, the authors found empirically that the values m_1 and m_2 that cancel respectively the asymptotic covariances

[1] https://github.com/TheoMoins/ExtremesPyMC

$\text{ACov}(\mu, \sigma)$ and $\text{ACov}(\sigma, \xi)$ are such that any $m \in [m_1, m_2]$ improves the MCMC convergence. Approximations of m_1 and m_2 are then given as functions of ξ, and therefore a preliminary estimation of ξ (typically using maximum likelihood estimation) is required to obtain $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$, and to choose a value in this interval before running an MCMC with the right choice of m. Despite a significant improvement of the convergence, this method has several limitations. First, a preliminary estimation of the shape parameter ξ is required, before computing $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$ and choosing a value in the corresponding interval, which adds complexity and computational burden on the overall framework. Moreover, it also affects the accuracy of orthogonalization, as the expressions of m_1 and m_2 are found empirically, then are approximated by $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$, and finally computed at $\hat{\xi}$ which adds a new source of uncertainty. One way to lighten the method would be to suggest a simpler choice of m, for example $m = n$, which leads to a satisfactory behaviour as noticed by Wadsworth et al. (2010). However, we show in Appendix A that this choice presents some flaws and does not bring any general guarantee of orthogonality.

Orthogonal parameterization More directly, there exists a parameterization of the Poisson process that leads to orthogonality. Suggested by Chavez-Demoulin and Davison (2005), it consists of the following change of variable:

$$(r, \nu, \xi) = \left(m \left(1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right)^{-1/\xi}, (1 + \xi)(\sigma + \xi(u - \mu)), \xi \right). \tag{7}$$

With this parameterization, the likelihood can be written as

$$L(\boldsymbol{x} \mid r, \nu, \xi) = e^{-r} \left(\frac{r}{m} \right)^n \left(\frac{\nu}{1 + \xi} \right)^{-n_u} \prod_{i=1}^{n_u} \left(1 + \frac{\xi(1 + \xi)}{\nu} (x_i - u) \right)^{-1-1/\xi}. \tag{8}$$

Under this form, we can directly see that r is orthogonal to ν and ξ, as the likelihood factorizes with respect to r and (ν, ξ). Parameter $r \geq 0$ represents the intensity of the Poisson process, which is the expected number of exceedances, while the two other ones can be seen as an orthogonal parameterization of the GPD distribution with scale $\tilde{\sigma}_u = \sigma + \xi(u - \mu)$ and shape ξ. Under this parameterization and if $\xi > -1/2$, the Fisher information matrix $\mathcal{I}(r, \nu, \xi)$ is finite, diagonal and can be written as

$$\mathcal{I}(r, \nu, \xi) = \text{diag} \left(\frac{1}{r}, \frac{r}{\nu^2(1 + 2\xi)}, \frac{r}{(1 + \xi)^2} \right),$$

where $\text{diag}(\boldsymbol{u})$ denotes the diagonal matrix with diagonal equal to vector $\boldsymbol{u}$. Calculation details are provided in Appendix B. Therefore, the orthogonal parameterization of Chavez-Demoulin and Davison (2005) is more adapted than the tuning of m since it directly yields the optimal solution sought by Sharkey and Tawn (2017). Moreover, it is obtained without recourse to any optimization procedure or approximation. Finally, by plugging the variables (r, ν) into Equation (4), we can show that the invariance property with respect to m holds for the three parameters, and so the parameterization is independent of the choice of m.

3 Priors invariant to reparameterization

In the case where no external information is available about the parameters, the choice of the prior distribution should be made with caution. Typically, the term "uninformative prior" or "objective

prior" can be misleading, as it refers to priors used when one does not have preliminary information, but the prior itself does contain information. As an example, a flat prior over the range of possible values does not seem to make any distinction, but depends on how it is parameterized: a uniform prior does not necessarily stay uniform after a change of parameters. This problem is all the more serious for our study which deals with reparametrization: for the Poisson process model, how to justify a uniform prior over $(\mu, \log\sigma, \xi)$, $(r, \log\frac{\nu}{1+\xi}, \xi)$, or something else? Even in the informative case, two experts that provide equivalent quantities on two parameterizations should expect the same result in the end. Here, we derive two priors that enjoy the property of being invariant with respect to reparameterization.

3.1 Jeffreys prior

Jeffreys prior (Jeffreys, 1946) is built with the aim of invariance: if $\mathcal{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix associated with parameters $\boldsymbol{\theta}$, it is defined as

$$p_{\text{J}}(\boldsymbol{\theta}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\theta})}.$$

Under this prior, one can show that a reparameterization $\boldsymbol{\phi} = h(\boldsymbol{\theta})$ yields $p_{\text{J}}(\boldsymbol{\phi}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\phi})}$. This prior is computed for the GPD by Castellanos and Cabras (2007) and for the GEV under a modified version where $p_J(\mu, \sigma, \xi) \propto \sqrt{\det \mathcal{I}(\sigma, \xi)}$ by (Kotz and Nadarajah, 2000). Up to our knowledge, Jeffreys prior has never been computed for the Poisson process characterization of extremes. Nevertheless, the orthogonalization done in Equation (7) directly provides Jeffreys prior with respect to (r, ν, ξ):

Proposition 1 *Jeffreys prior associated with a Poisson process for extremes with parameters (r, ν, ξ) from Equation (7) exists provided $\xi > -1/2$, and can be written as*

$$p_{\text{J}}(r, \nu, \xi) \propto \frac{r^{1/2}}{\nu(1+\xi)(1+2\xi)^{1/2}}. \tag{9}$$

Moreover, the invariance to reparameterization property provides directly the expression of Jeffreys prior on (μ, σ, ξ).

Corollary 1 *Jeffreys prior associated with a Poisson process for extremes with original parameters (μ, σ, ξ) exists provided $\xi > -1/2$, and can be written as*

$$p_J(\mu, \sigma, \xi) \propto \frac{\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{3}{2\xi}-1}}{\sigma^2(1+\xi)(1+2\xi)^{1/2}}. \tag{10}$$

Note that this prior, similarly to the uniform one, is improper in the sense that the integral over the range of parameters is infinite. Consequently, it is necessary to check whether the posterior is proper or not to be able to use it. Castellanos and Cabras (2007) shows that the posterior is proper when using Jeffreys prior in the GPD case, while Northrop and Attalides (2016) shows that it is never the case with GEV likelihood. For the Poisson process, we show the following result:

Proposition 2 *Jeffreys prior for a Poisson process for extremes yields a proper posterior distribution, as soon as $\xi > -1/2$.*

A proof is provided in Appendix B.

3.2 Penalized complexity prior for the shape parameter

The shape parameter ξ plays a crucial role in the estimation, as it tunes the heaviness of the tail distribution: it is heavy if $\xi > 0$, light if $\xi = 0$ and finite (*i.e.* with a finite right end-point) if $\xi < 0$. The case $\xi = 0$ can be seen as a simpler model with an exponential decrease of the survival function, where the GPD cdf in Equation (2) simplifies to an exponential distribution. This concentration of an entire maximum domain of attraction at a single value of ξ complicates the study, as it is for example difficult to distinguish heavy tails with low ξ and light tails (Stephenson and Tawn, 2004). However, this change of regime can have significant consequences when it comes to extrapolation. It should also be noted that a vast majority of datasets have distribution with $|\xi|$ value less than 1. It is therefore natural, even in a non-informative framework, to favor the case $\xi = 0$ and penalize high values of $|\xi|$. One way to do this is to use penalized complexity (PC) priors (Simpson et al., 2017): the idea is to consider a prior that penalizes exponentially the distance between a model $p_\xi := p(\cdot \mid \xi)$ with a given ξ and the baseline p_0 with $\xi = 0$. The general formula is given by

$$p_{\mathrm{PC}}(\xi \mid \lambda) = \lambda \exp(-\lambda d(\xi)) \left| \frac{\partial d(\xi)}{\partial \xi} \right|,$$

with $\lambda > 0$, $d(\xi) = \sqrt{2\mathrm{KL}\left(p_\xi || p_0\right)}$ and $\mathrm{KL}(p_\xi||p_0)$ the Kullback–Leibler divergence between p_ξ and p_0: $\mathrm{KL}(p_\xi||p_0) = \int p_\xi(x) \log\left(p_\xi(x)/p_0(x)\right) \mathrm{d}x$. Parameter λ acts as a scaling parameter and controls the range of acceptable values for ξ. This prior has the advantage of being proper and invariant to reparameterization on ξ. The computation with GPD has already been done by Opitz et al. (2018) for the case $\xi \geq 0$: the authors show that $d(\xi)$ is finite only if $\xi < 1$, and is given by $d(\xi) = \sqrt{2}\xi/\sqrt{1-\xi}$ for $0 \leq \xi < 1$. Then, they show that it can be approximated by an exponential distribution on ξ in the case $\xi \to 0$, when λ can be taken large and favor sufficiently $\xi = 0$. A first observation is that routine calculations extend definition to negative values of ξ, and for the Poisson process characterization where the density of observation is also GPD:

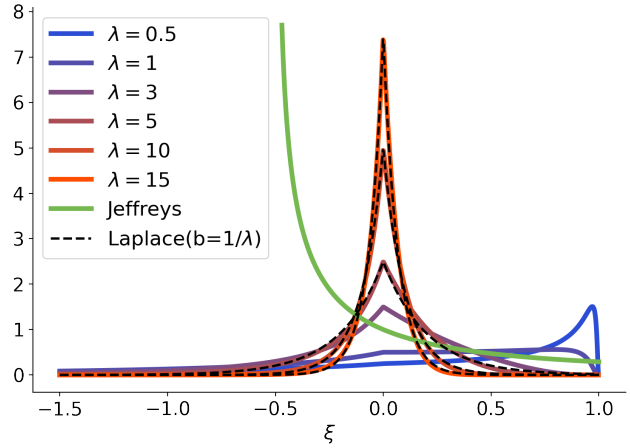
Proposition 3 *PC prior associated with a Poisson process for extremes exists for any $\xi < 1$ and can be written as*

$$p_{\mathrm{PC}}(\xi \mid \lambda) = \frac{\lambda}{2} \left(\frac{1 - \xi/2}{(1-\xi)^{3/2}} \right) \exp\left(-\lambda \frac{|\xi|}{\sqrt{1-\xi}} \right). \tag{11}$$

This prior is plotted for several values of λ in Figure 1. Similarly to the observation of Opitz et al. (2018), this prior is very similar to a Laplace$(0, 1/\lambda)$ when λ is sufficiently high for the peak at 0 to dominate over the endpoint at 1. In the case when zero is favoured with a high λ and $\xi \neq 0$, the estimation may be altered compared to the uninformative case: see Appendix C.4 for an analysis on simulated data. For the two other parameters, one can consider the Jeffreys' rule on (r, ν) in order to obtain a non-informative approach for (r, ν) while keeping invariance to reparameterization property (ξ is therefore considered a priori independent of (r, ν)). Looking at the Fisher information matrix in Equation (8), we obtain $p_{\mathrm{J}}(r, \nu) \propto 1/\nu$. Similarly to Jeffreys prior in Section 3.1, the resulting prior is improper but we can show the following proposition:

Proposition 4 *The prior defined as $p(r, \nu, \xi) \propto p_{\mathrm{PC}}(\xi) p_{\mathrm{J}}(r, \nu) \propto p_{\mathrm{PC}}(\xi)/\nu$ for the Poisson process for extremes yields a proper posterior distribution.*

The proof, detailed in Appendix B, relies on a result of Northrop and Attalides (2016). Note that this result still holds if $p_{\mathrm{PC}}(\xi)$ is replaced by its Laplace approximation.



λ	$\mathrm{I}_{\mathrm{PC}}(0.95 \mid \lambda)$	$\mathrm{I}_{\mathrm{Lap}}(0.95 \mid \lambda)$
0.5	[-36.8, 0.97]	[-6.0, 6.0]
1	[-9.88, 0.90]	[-2.99, 2.99]
3	[-1.61, 0.61]	[-0.99, 0.99]
5	[-0.80, 0.44]	[-0.59, 0.59]
10	[-0.34, 0.25]	[-0.29, 0.29]
15	[-0.22, 0.18]	[-0.19, 0.19]

Figure 1: Left panel: examples of PC priors $p_{\mathrm{PC}}(\cdot \mid \lambda)$ with λ from 0.5 (blue curve) to 15 (red curve), and Jeffreys prior (green curve) represented for fixed values of (μ, σ). The black dashed lines represent Laplace distributions with scale parameter equal to $1/\lambda$, for $\lambda \in \{5, 10, 15\}$. Note that Laplace distributions $p_{\mathcal{L}}(\cdot \mid 1/\lambda)$ approximate well $p_{\mathrm{PC}}(\cdot \mid \lambda)$ when $\lambda \geq 10$. Right panel: corresponding intervals at 95% for PC and Laplace priors, resp. $\mathrm{I}_{\mathrm{PC}}(0.95 \mid \lambda)$ and $\mathrm{I}_{\mathrm{Lap}}(0.95 \mid \lambda)$.

4 Experiments

In this section, we illustrate the benefits of the orthogonal reparameterization on simulations and a real environmental dataset. All experiments are done using PyMC3 library (Salvatier et al., 2016), and the corresponding code is available online (link in the Introduction).

4.1 Simulations with the Poisson process model

Data generation We start by comparing the different parameterizations on exceedances generated with the Poisson process model described in Section 1.1. For a given value of (μ, σ, ξ) and hyperparameters (u, m), the data generation proceeds in two steps: first, a number of events n is simulated using a Poisson distribution with density $\Lambda(I_u)$ as defined in Section (1.1). Then, for each point $i \in \{1, \ldots, n\}$, the position x_i knowing that $x_i \in I_u$ is sampled from a GPD with parameters $(u, \tilde{\sigma}, \xi)$, with $\tilde{\sigma} = \sigma + \xi(u - \mu)$. An example with $(m, u, \mu, \sigma, \xi) = (40, 30, 50, 15, -0.25)$ is detailed here, which leads to an expected number of observations $\Lambda(I_u) \approx 126$.

Experimental setup For MCMC hyper-parameters such as number of chains, burn-in period per chain or initial values, we keep the default values suggested in the PyMC3 library: in particular, the number of chains is equal to $\max\{n_c, 2\}$ with n_c the number of cores (in our case $n_c = 4$), and the burn-in period is set to 1 000. In addition to these choices, this library offers the possibility to choose among different sampling methods, such as the traditional Metropolis–Hastings algorithm, but also more modern MCMC algorithms like Hamiltonian Monte Carlo (HMC, Neal, 2011), or the No-U-Turn sampler (NUTS, Hoffman and Gelman, 2014) which is the default choice in PyMC3. We choose to compare the different reparameterizations on 1 000 Metropolis–Hastings draws (after

burn-in), and the behaviour on NUTS is also investigated in Appendix C.1. Finally, the prior we choose for all our configurations is Jeffreys prior, computed in Section 3.1, but experiments have shown similar results with the PC prior of Section 3.2.

Convergence diagnostic Our aim is to discriminate the different parameterizations according to the rate of convergence of the MCMC chains to their target. More precisely, for a finite number of iterations, two properties should be verified: stationarity, to check if the chains are still in an exploration phase, and mixing, which corresponds in practice to checking if the whole parameter space has been explored. Different indicators exist to quantify these properties. First, autocorrelation plots as a function of lag measure how good the posterior approximation is, as the dependence between the elements of the chains reduce the effective information available for inference. To measure this, common practice relies the effective sample size, defined as $\text{ESS} = MN(1 + 2\sum_{t=1}^{\infty} \rho_t)^{-1}$, with M the number of chains of size N, and ρ_t the autocorrelation at lag t. It corresponds to the equivalent number of independent draws for estimation, and so quantifies the amount of effective data for estimation. More details can be found in Gelman et al. (2013, Section 11.5). Here, the evolution of ESS with the number of draws for each configuration is reported. To complete the diagnostic, the potential scale reduction factor (commonly denoted by $\hat{R}$) also aims at bringing an indication about the state of convergence by computing the ratio of two estimators of the posterior variance (Gelman and Rubin, 1992). Generally $\hat{R} \geq 1$, and if it is greater than a given threshold, a convergence issue is raised. We use here a refinement of $\hat{R}$ named $\hat{R}_\infty$ (Moins et al., 2022), based on a local version $\hat{R}(x)$ which aims at ensuring the convergence at a given quantile x of the distribution. Then, $\hat{R}_\infty$ is defined as the supremum of the different $\hat{R}(x)$ values: $\hat{R}_\infty := \sup_{x \in \mathbb{R}} \hat{R}(x)$. This scalar summary corresponds to considering the value of $\hat{R}(x)$ associated with the worse quantile approximation by the MCMC chains.

Results Results are reported in Figure 2. Three parameterizations are compared for MCMC efficiency: one with (μ, σ, ξ) and no changes on u and m as given in Equation (5), one with the triplet (μ, σ, ξ) that corresponds to a choice of m suggested by Sharkey and Tawn (2017) (see Section 2), and the orthogonal parameterization (r, ν, ξ) of Equation (7). In order to compare the same quantities, all convergence diagnostics are computed on the original parameterization (μ, σ, ξ) with the original value of m, so after a transformation of the chains for the second and third parameterizations. In view of Figure 2, we can confirm that the orthogonal parameterization behaves best in the case $\xi < 0$: the parameters have the lowest autocorrelations in the chains, the lowest value of $\hat{R}(x)$ for almost all x, and this parameterization is the only one which satisfies the recommendation of having an ESS ≥ 400 for estimation (Gelman et al., 2013). Conversely, the two other parameterizations seem to suffer from a lack of convergence, and in particular the one suggested by Sharkey and Tawn (2017). Some intuitions about these issues can be found in Appendix A together with additional experiments in the cases $\xi > 0$ and $\xi = 0$ in Appendix C.2 leading to similar results. We also refer to Appendix C.3 for experiments in the GPD case. As a conclusion, the orthogonal parameterization is effective in the three maximum domains of attraction, for both Poisson and GPD models.

4.2 Case study on river flow data

We apply our framework on daily measurements of the Garonne river flow (France), from 1915 to 2013, which represent a total of 36 160 observations.

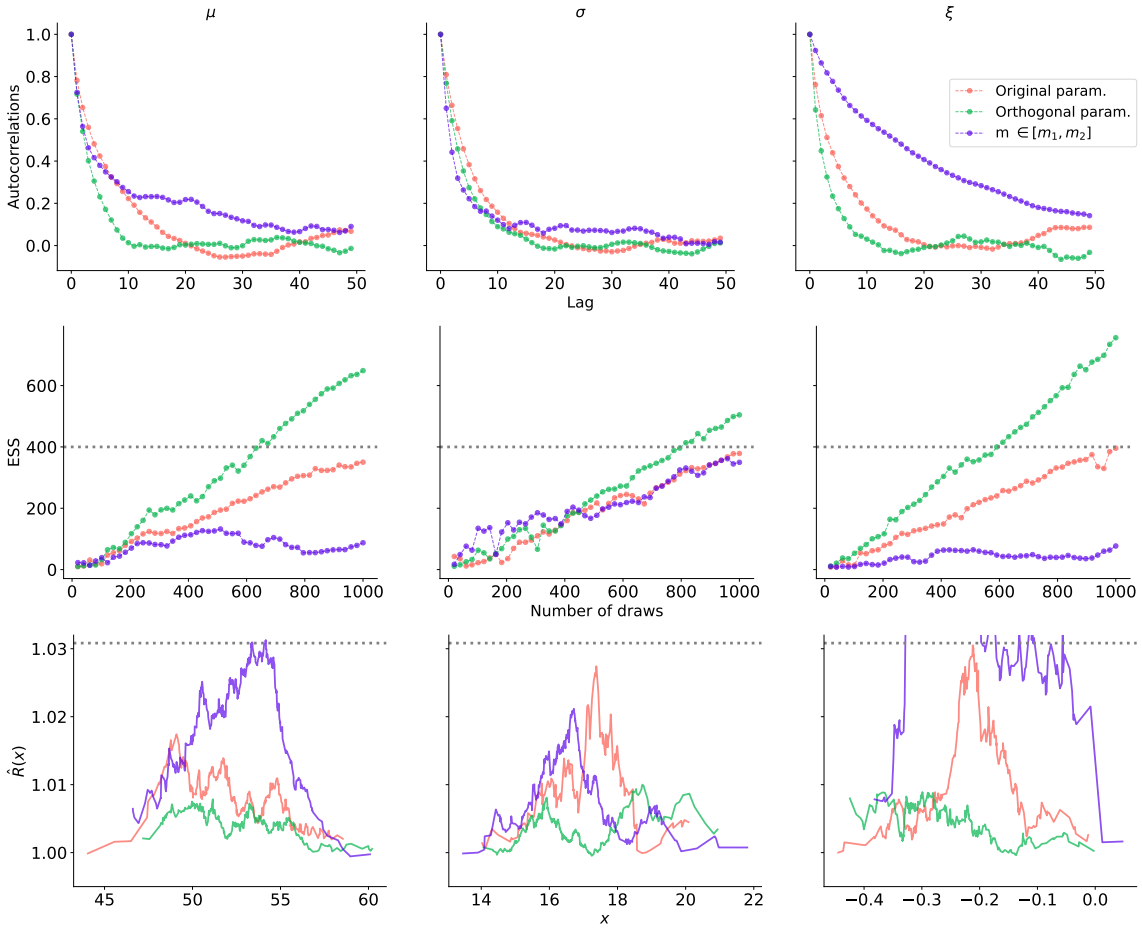


Figure 2: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi < 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for three different parameterizations: the original one (in red), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in blue), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.031 (see Moins et al., 2022).

Preprocessing Before selecting a threshold and running an MCMC algorithm, some common preprocessing steps on daily environmental data are required: first because of seasonality, we consider only the rainy season from December to May, which reduces the number of observations to 18 043. Also, the observations are not independent and an auto-correlation plot suggests a three-day correlation in measurements. Therefore, clusters of exceedances of parameters $r = 3$ days are considered here, which means that two exceedances that occurred in less than three days are merged as one observation (the largest one in the cluster). Previous EDF studies (see for instance Chapter 4 of Albert, 2018) agree with traditional threshold elicitation methods (see Coles, 2001) to consider a threshold of $u = 2\,000$ m^3/s for estimation. In the end, we obtain a total of $n = 182$ clusters of exceedances which are represented in Figure 3.

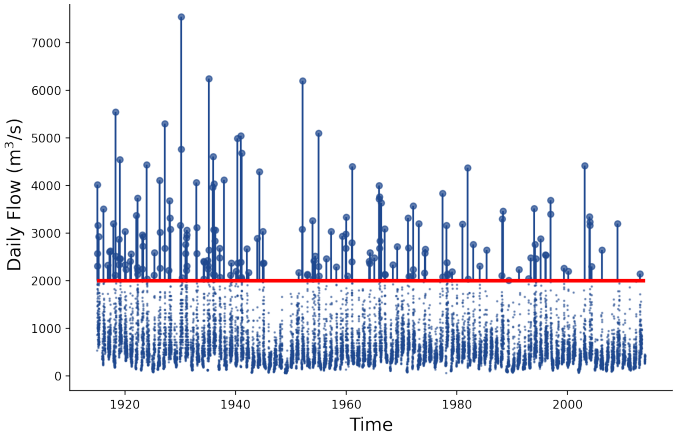


Figure 3: Plot of $n = 182$ exceedances of the Garonne river flow between 1915 and 2013 above the threshold $u = 2\,000$ (represented in red).

Return level estimation We are interested in estimating the T-year return level ℓ_T, which is exceeded on average once every T years. This is obtained by solving the equation $G(\ell_T \mid \mu, \sigma, \xi) = 1 - 1/T$, with G the GEV cdf defined in Equation (1):

$$\ell_T = \mu - \frac{\sigma}{\xi}\left(1 - (-\log(1 - 1/T))^{-\xi}\right). \tag{12}$$

Here, as the data span 99 years, we fix $m = 99$ in order to obtain parameters associated with annual maxima. The same setup as in Section 4.1 is then run with 5 000 draws from Metropolis–Hastings algorithm with the orthogonal parameterization. Convergence diagnostic values are reported in Figure 4 and show no evidence of convergence issue, along with a very satisfactory effective sample size for estimation (final values can be found in Table 1 along with $\hat{R}_\infty$ for each parameters). Results of posterior summaries for (μ, σ, ξ) are reported in Table 1: looking at the posterior for ξ, the three maximum domains of attraction cannot be excluded, although the credible interval (CI) at 95% is tight around zero. This may suggest that $\xi = 0$ and an exponential decrease of the survival function. Return levels for annual maxima are represented in the left panel of Figure 5, and show that the model seems to fit the data correctly. This curves are obtained by computing the mean and 2.5%/97.5% quantiles on the posterior distribution of ℓ_T for any given return period T. This

	Post. Mean	Post. SD	95%-CI	ESS	$\hat{R}_\infty$
μ	2 560.8	84.1	[2 409.8, 2 724.1]	3 473	≈ 1.0
σ	919.6	73.2	[787.2, 1 063.3]	2 709	≈ 1.0
ξ	0.015	0.077	[−0.120, 0.164]	2 702	≈ 1.0

Table 1: Posterior summaries (mean, standard deviation (SD), credible interval (CI) at 95%) and convergence diagnostics (ESS and $\hat{R}_\infty$) for (μ, σ, ξ) associated with annual maxima ($m = 99$).

is more accurate than the version where pointwise posterior quantities of (μ, σ, ξ) are plugged in Equation (12) (see Jonathan et al., 2021, for a comparison). The obtained posterior mean of ℓ_T, is 6 949 m^3/s for the 100-year level and 9 266 m^3/s for the 1 000-year one. These results corroborate a study conducted in Albert et al. (2020), as their estimated value of 10 000 m^3/s for the 1 000-year return level belongs to the credible interval in Figure 5.

Prior influence on the return level estimation uncertainty Looking at the posterior distribution for ξ, one can reasonably make the assumption that $\xi = 0$ and therefore assume an exponential decrease for the survival function of the river flow. In this case, the remaining location parameter μ and scale parameter σ can be estimated with fixed $\xi = 0$. The resulting posterior summaries are very similar to the ones of Table 1. As a result, the return level curves with posterior mean parameters (see Figure 5) are very similar in both cases. However, as the uncertainty on the shape parameter is excluded when fixing $\xi = 0$, the return levels credible intervals change drastically and become very concentrated around means, as shown in the right panel of Figure 5. In fact, this reflects that most of the uncertainty on the estimated return level is due to the estimation of the shape parameter, and so knowing its value facilitates greatly the extrapolation. PC priors allow to navigate between these two extreme cases thanks to the hyperparameter λ. Looking at the left panel of Figure 6, it appears that the return level curves associated with posterior means are not affected by those differences of priors. However, the larger λ, the more information is added about the closeness of ξ to zero, and the smaller the length of the credible interval (note however that this does not give any guarantee on the estimation bias). This behaviour is illustrated on the right panel of Figure 6: if we denote by $\ell_T^{(\mathrm{m})}$, $\ell_T^{(2.5\%)}$, and $\ell_T^{(97.5\%)}$ respectively the posterior mean, and the posterior quantiles at 2.5% and 97.5% of the return level, then the right plot in Figure 6 displays the length of the credible interval for the return level estimation, relatively to the estimator $\ell_T^{(\mathrm{m})}$: $(\ell_T^{(97.5\%)} - \ell_T^{(2.5\%)})/\ell_T^{(\mathrm{m})}$. This ratio is expected to grow with T, as the uncertainty increases in the tail. When $\lambda = 1$, this growth is similar to the one associated with Jeffreys prior, which can be seen as a noninformative case. For example, we can see that the size of the credible interval is already greater than the posterior estimation for the 1 000-year return level (ratio greater than one). Then using $\lambda = 10$, which corresponds to a confidence of 95% of having ξ between -0.3 and 0.3 with the Laplace approximation (see the table in Figure 1), reduces by approximately 20% the size of the credible interval for $T = 1\,000$. The length when ξ is fixed at zero is drastically lower than in the other cases, even those concerning PC priors with large λ values. In our case where this dataset has already been studied in the past and a value very close to zero was expected, a choice of $\lambda = 10$ seems reasonable for the PC prior.

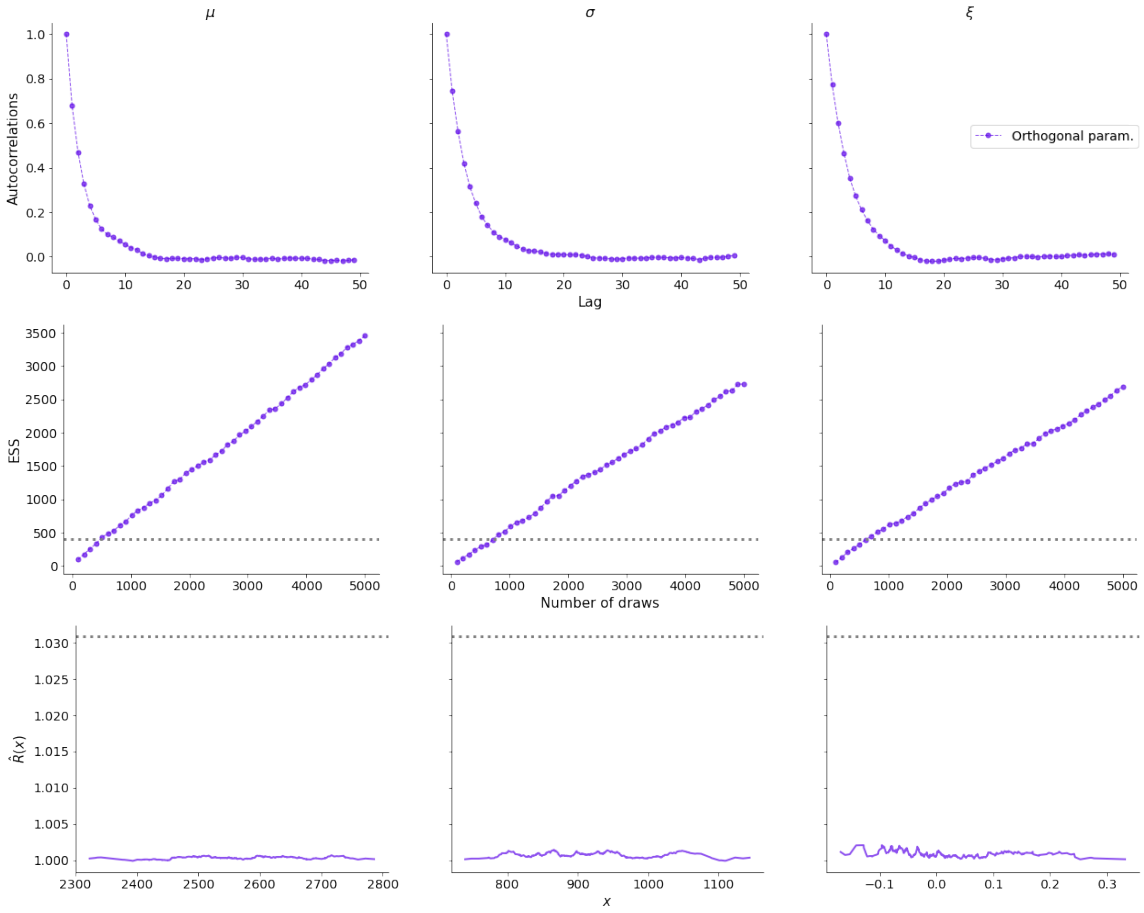


Figure 4: Convergence diagnostic plots for Garonne river flow data, after 5 000 Metropolis–Hastings draws and a burn-in of 1 000. Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.031 (see Moins et al., 2022).

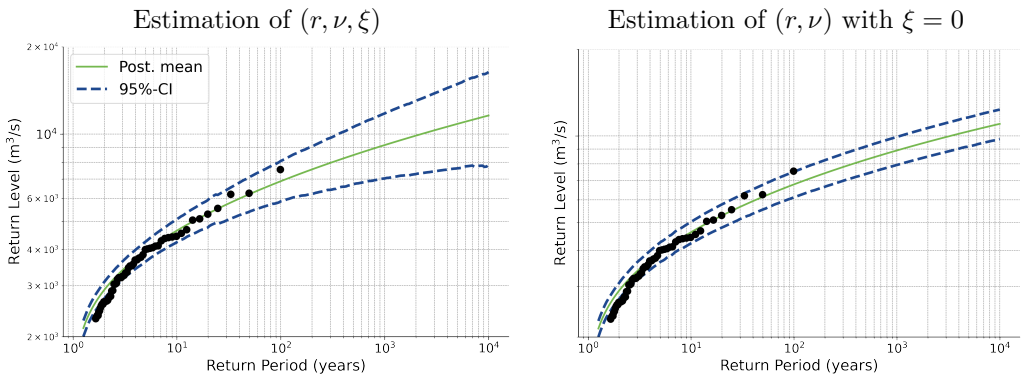


Figure 5: Return levels for annual maxima of Garonne flow data. Full green curves correspond to return levels obtained with posterior mean parameters, and the dashed ones to the bounds of the 95% credible interval (CI). On the left, all three parameters (r, ν, ξ) are estimated, while on the right, only (r, ν) are estimated with the assumption that $\xi = 0$. The black points represent the observed annual maxima.

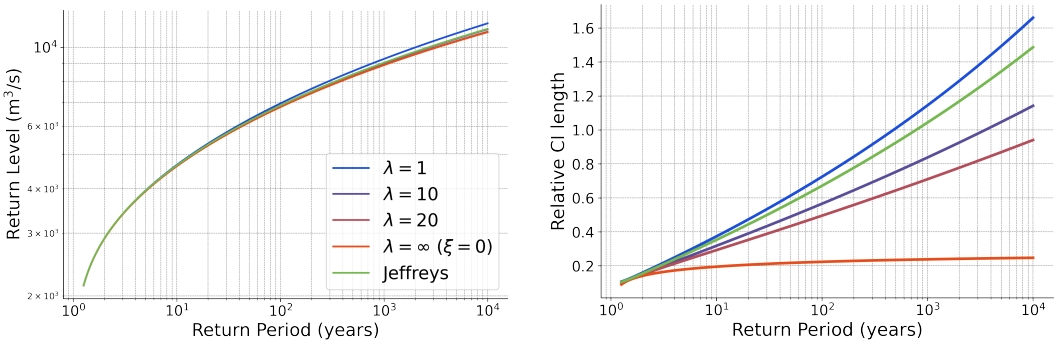


Figure 6: Comparison of return levels with different priors as functions of return period (log scale). On the left: return levels with posterior mean parameters. On the right: return level credible interval (CI) length relative to the point estimate (in %).

5 Conclusion

In this paper we demonstrate the benefits of using an orthogonal parameterization in the sense of Jeffreys (1961) for Bayesian inference of extreme value models. First, orthogonal parameters facilitate the convergence of MCMC algorithms such as Metropolis–Hastings or NUTS (Section 2 and Appendix A). This improvement is "for free" in the sense that it is obtained at no extra computational cost, except a simple change of variable if one interest lies in the original parameters (μ, σ, ξ). This conclusion is confirmed by several convergence diagnostics on simulations in the three maximum domains of attraction (Section 4.1 and Appendix C).

Secondly, the orthogonal parameterization also facilitates the computation of Jeffreys prior (Section 3.1): we show that this uninformative prior is defined for $\xi > -1/2$ and is improper, but leads to a proper posterior. Posterior propriety is a necessary condition for using this prior in practice, when no external information is available. However, this uninformative case is actually far from the reality of most of the applications: in practice, even without any expert information, a shape parameter in the range of -1 and 1 already includes a vast majority of the distributions arising in natural phenomena. Therefore as an alternative, a PC prior on ξ can be used instead and allows to control the prior knowledge one wants to include on ξ (Section 3.2). In particular, it allows to penalize the values of ξ that move away from 0, and navigate between the uninformative case and the deterministic one where $\xi = 0$. In addition to its flexibility, this prior enjoys the same advantages as Jeffreys prior: invariance to reparameterization and posterior propriety. Additionally, it can be defined without any restriction for ξ if one uses the approximation by a Laplace distribution (otherwise, $\xi < 1$). This prior information on ξ impacts the posterior uncertainty around the return level estimation. By applying our framework on river flow data (Section 4.2), we showed that the length of the credible interval for the return level can be significantly reduced by adding prior information of ξ, until the extreme case where we assume a light tail ($\xi = 0$). However, the uncertainty around the return level can be quantified differently, by using the quantiles of the posterior predictive distribution defined in (6), see Fawcett and Green (2018) for a comparison. In future work, it would be interesting to also study the influence of the prior on the posterior predictive return levels.

References

Albert, C. (2018). *Estimation des limites d'extrapolation par les lois de valeurs extrêmes. Application à des données environnementales.* PhD thesis (in French), Université Grenoble Alpes.

Albert, C., Dutfoy, A., Gardes, L., and Girard, S. (2020). An extreme quantile estimator for the log-generalized Weibull-tail model. *Econometrics and Statistics*, 13:137–174.

Betancourt, M. (2019). Incomplete Reparameterizations and Equivalent Metrics. arXiv:1910.09407.

Betancourt, M. and Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4.

Bousquet, N. (2021). *Extreme Value Theory with Applications to Natural Hazards: From Statistical Theory to Industrial Practice.* Springer, Cham.

Browne, W. J., Steele, F., Golalizadeh, M., and Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models

with applications to discrete time survival models. *Journal of the Royal Statistical Society: Series A*, 172(3):579–598.

Castellanos, M. E. and Cabras, S. (2007). A default Bayesian procedure for the generalized Pareto distribution. *Journal of Statistical Planning and Inference*, 137(2):473–483.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C*, 54(1):207–222.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag, London.

Coles, S. G. and Powell, E. A. (1996). Bayesian Methods in Extreme Value Modelling: A Review and New Developments. *International Statistical Review*, 64(1):119–136.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B*, 49(1):1–18.

Fawcett, L. and Green, A. C. (2018). Bayesian posterior predictive return levels for environmental extremes. *Stochastic Environmental Research and Risk Assessment*, 32(8):2233–2252.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). Efficient parametrizations for generalized linear mixed models. *Bayesian Statistics*, 5:48–74.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. arXiv:2011.01808.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Gilleland, E. and Katz, R. W. (2016). extRemes 2.0: an extreme value analysis package in R. *Journal of Statistical Software*, 72:1–39.

Haan, L. and Ferreira, A. (2006). *Extreme value theory: an introduction*, volume 21. Springer.

Hills, S. E. and Smith, A. F. (1992). Parameterization issues in Bayesian inference. *Bayesian Statistics*, 4:227–246.

Hoffman, M. D. and Gelman, A. (2014). No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Huzurbazar, V. S. (1950). Probability distributions and orthogonal parameters. *Mathematical Proceedings of the Cambridge Philosophical Society*, 46(2):281–284.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London: Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Jeffreys, H. (1961). *The Theory of Probability*. Oxford, third edition.

Jonathan, P., Randell, D., Wadsworth, J., and Tawn, J. (2021). Uncertainties in return values from extreme value analysis of peaks over threshold using the generalised pareto distribution. *Ocean Engineering*, 220:107725.

Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press.

Leadbetter, M., Lindgren, G., and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer, Germany.

Moins, T., Arbel, J., Dutfoy, A., and Girard, S. (2022). On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic. arXiv:2205.06694.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2.

Northrop, P. J. and Attalides, N. (2016). Posterior propriety in Bayesian extreme value analyses using reference priors. *Statistica Sinica*, 26(2):721–743.

Opitz, T., Huser, R., Bakka, H., and Rue, H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes*, 21(3):441–462.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. *Bayesian Statistics*, 7:307–326.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131.

Roberts, G. O. and Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society: Series B*, 56(2):377–384.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B*, 59:291–317.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55.

Sharkey, P. and Tawn, J. A. (2017). A Poisson process reparameterisation for Bayesian inference for extremes. *Extremes*, 20(2):239–263.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.

Stephenson, A. (2016). Extreme Value Modeling and Risk Analysis: Methods and Applications.

Stephenson, A. and Tawn, J. (2004). Bayesian inference for extremes: Accounting for the three extremal types. *Extremes*, 7(4):291–307.

Tibshirani, R. and Wasserman, L. (1994). Some aspects of the reparametrization of statistical models. *Canadian Journal of Statistics*, 22:163–173.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718.

Wadsworth, J. L., Tawn, J. A., and Jonathan, P. (2010). Accounting for choice of measurement scale in extreme value modeling. *Annals of Applied Statistics*, 4(3):1558–1578.

Woutersen, T. (2011). Consistent estimation and orthogonality. *Advances in Econometrics*, 27(1):155–178.

Appendix A Approaching orthogonality by choosing $m = n$

Sharkey and Tawn (2017) aims at choosing a value of the scaling factor m that minimises the off-diagonal terms of the asymptotic covariance matrix (that is the inverse Fisher information matrix), denoted by $\text{ACov} := \mathcal{I}^{-1}(\mu, \sigma, \xi)$. Those terms exist only if $\xi > -1/2$ (see Proposition 2 and its proof in Appendix B) and can be written as functions of $x = -\frac{1}{\xi}\log\left\{1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right\}_+$, σ, and ξ as:

$$\begin{aligned}
\text{ACov}_{\mu,\sigma} &= \frac{\sigma^2}{m\xi^2} e^x \left(\xi^3 + (1+\xi)(1+2\xi+\xi(1+\xi)x^2 - (1+3\xi)x + e^{-\xi x}(1+2\xi)(x-1))\right), \\
\text{ACov}_{\mu,\xi} &= \frac{\sigma}{m\xi^2} e^x (1+\xi)\left(\xi(1+\xi)x - (1+2\xi)(1-e^{-\xi x})\right), \\
\text{ACov}_{\sigma,\xi} &= \frac{\sigma}{m} e^x (1+\xi)\left((1+\xi)x - 1\right).
\end{aligned}$$

Denoting by $\rho_{.,.}$ the asymptotic correlation between two out of the three parameters, the authors note that a range of values may also work for m between m_1 and m_2, where

$$m_1 = \underset{m}{\text{argmin}}\{|\rho_{\mu,\sigma}| + |\rho_{\mu,\xi}|\} \text{ and } m_2 = \underset{m}{\text{argmin}}\{|\rho_{\mu,\sigma}| + |\rho_{\sigma,\xi}|\}.$$

They also find on their experiments that m_1 cancels $\rho_{\mu,\sigma}$, and that m_2 cancels $\rho_{\sigma,\xi}$. A numerical method is used in Sharkey and Tawn (2017) to approximate m_1 and m_2 as functions of ξ. Therefore, this approach requires to study the roots x_1 of $\text{ACov}_{\sigma,\xi}$ and x_2 of $\text{ACov}_{\mu,\sigma}$ to respectively deduce $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$. Without any approximation, we directly have $x_1 = 1/(1+\xi)$ as the unique root for $\text{ACov}_{\sigma,\xi}$. Moreover, as $\xi > -1/2$, we have $x_1 > 0$, which motivates us to study the sign of the root x_2 for $\text{ACov}_{\mu,\sigma}$. Indeed, if x_2 is unique and $x_2 < 0$, then the choice $x = 0$ which cancels the third asymptotic covariance $\text{ACov}_{\mu,\xi}$ will always be reasonable as it will stay in the targeted interval, between the two other roots. In addition, $x = 0$ corresponds to the choice $m = r$ (which in

practice translates into $m = n$), and is a simple choice as it does not require any estimation of ξ. The interest of the choice $m = n$ has already been mentioned in Wadsworth et al. (2010) to improve the mixing property of the chain. Unfortunately, a study of function $x \mapsto \mathrm{ACov}_{\mu,\sigma}(x)$ shows that the properties of uniqueness and positivity of x_2 are only valid in the case where $\xi > 0$. In that case, studies of Wadsworth et al. (2010) and Sharkey and Tawn (2017) corroborate the choice of $m = n$. However, it is not the case anymore when $-1/2 < \xi < 0$. It can be shown that x_2 is not negative here, and worse, may not be unique. This can be seen as contraindications for frameworks that aims at reducing the three asymptotic covariances at the same time by tuning the scaling factor m.

Appendix B Proofs associated with prior computations

Proof of Proposition 1 The log-likelihood l using the (r, ν, ξ) parameterization of Equation (7) can be written as:

$$l(r,\nu,\xi \mid \boldsymbol{x}, n) = -r + n\log\left(\frac{r}{m}\right) - n\log(\nu) + n\log(1+\xi) - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+.$$

Under this form, we can directly see that r is orthogonal to ν and ξ. The second derivatives are given by

$$\begin{aligned}
\frac{\partial^2 l}{\partial r^2} &= -\frac{n}{r^2},\\
\frac{\partial^2 l}{\partial r \partial \nu} &= 0,\\
\frac{\partial^2 l}{\partial r \partial \xi} &= 0,\\
\frac{\partial^2 l}{\partial \nu^2} &= \frac{n}{\nu^2} + \frac{\xi(1+\xi)^3}{\nu^4}\sum_{i=1}^{n}\frac{(x_i-u)^2}{\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+^2} - \frac{2(1+\xi)^2}{\nu^3}\sum_{i=1}^{n}\frac{(x_i-u)}{\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+},\\
\frac{\partial^2 l}{\partial \nu \partial \xi} &= \frac{(1+2\xi)(1+\xi)^2}{\nu^3}\sum_{i=1}^{n}\frac{(x_i-u)^2}{\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+^2} - \frac{2(1+\xi)}{\nu^2}\sum_{i=1}^{n}\frac{(x_i-u)}{\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+},\\
\frac{\partial^2 l}{\partial \xi^2} &= -\frac{n}{(1+\xi)^2} + \frac{(1+2\xi)^2(1+\xi)}{\xi\nu^2}\sum_{i=1}^{n}\frac{(x_i-u)^2}{\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+^2}\\
&\quad + \frac{2(1+\xi-\xi^2)}{\xi^2\nu}\sum_{i=1}^{n}\frac{(x_i-u)}{\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+} - \frac{2}{\xi^3}\sum_{i=1}^{n}\log\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+.
\end{aligned}$$

For the expectations, as we observe a Poisson process, the information is contained in the number n of observed points (we write N the corresponding random variable) and the position of jumping events x_i (we write X_i the corresponding random variable, with the same distribution as X). Here, N is distributed according to a Poisson distribution with parameter r, and $X - u$ is a GPD random variable with parameters $(\frac{\nu}{1+\xi}, \xi)$. For example, deriving the following expectation is the cornerstone

to obtain different terms of Fisher information matrix:

$$\begin{aligned}\mathbb{E}_{N,X}\left[\sum_{i=1}^{N}\frac{(X_i-u)^2}{\left\{1+\frac{\xi(1+\xi)}{\nu}(X_i-u)\right\}_+^2}\right] &= \mathbb{E}_N\left[\mathbb{E}_{X|N}\left[\sum_{i=1}^{N}\frac{(X_i-u)^2}{\left\{1+\frac{\xi(1+\xi)}{\nu}(X_i-u)\right\}_+^2}\right]\right]\\ &= \mathbb{E}_N\left[N\right]\mathbb{E}_{X|N}\left[\frac{(X-u)^2}{\left\{1+\frac{\xi(1+\xi)}{\nu}(X-u)\right\}_+^2}\right]\\ &= r\frac{1+\xi}{\nu}\int_u^{+\infty}(x-u)^2\left\{1+\frac{\xi(1+\xi)}{\nu}(x-u)\right\}_+^{-1/\xi-3}\mathrm{d}x.\end{aligned}$$

The above integral exists provided $\xi > -1/2$ and we obtain

$$\mathbb{E}_{N,X}\left[\sum_{i=1}^{N}\frac{(X_i-u)^2}{\left\{1+\frac{\xi(1+\xi)}{\nu}(X_i-u)\right\}_+^2}\right] = \frac{2r\nu^2}{(1+\xi)^3(1+2\xi)}.$$

Similarly, the remaining expected values can be written as

$$\begin{aligned}\mathbb{E}_{N,X}\left[\sum_{i=1}^{N}\frac{(X_i-u)}{\left(1+\frac{\xi(1+\xi)}{\nu}(X_i-u)\right)}\right] &= \frac{r\nu}{(1+\xi)^2},\\ \mathbb{E}_{N,X}\left[\sum_{i=1}^{N}\log\left(1+\frac{\xi(1+\xi)}{\nu}(X_i-u)\right)\right] &= r\xi.\end{aligned}$$

Plugging these values into the Fisher coefficients yields the result:

$$I(r,\nu,\xi) = \mathrm{diag}\left(\frac{1}{r}, \frac{r}{\nu^2(1+2\xi)}, \frac{r}{(1+\xi)^2}\right).$$

Proof of Proposition 2 Let us show that the following integral exists for any $n \geq 1$:

$$C_n = \iiint_{\mathcal{S}}\frac{r^{1/2}}{\nu(1+\xi)(1+2\xi)^{1/2}}e^{-r}\left(\frac{r}{m}\right)^n\left(\frac{\nu}{1+\xi}\right)^{-n}\prod_{i=1}^{n}\left(1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right)^{-1-\frac{1}{\xi}}\mathrm{d}r\mathrm{d}\nu\mathrm{d}\xi,$$

where $\mathcal{S}$ is the integration domain:

$$\mathcal{S} = \left\{(r,\nu,\xi)\in\mathbb{R}^3 \text{ s.t. } \xi > -\frac{1}{2},\, r>0,\, \nu \geq \{-\xi(1+\xi)((\max_i x_i)-u)\}_+\right\}.$$

Let us consider the case of one observation: $n = 1$, and $x_1 = x > u$. We have

$$
\begin{aligned}
C_1 &= \frac{1}{m}\int_{-1/2}^{+\infty}(1+2\xi)^{-1/2}\int_0^{+\infty} r^{3/2}e^{-r}\int_{\{-\xi(1+\xi)(x-u)\}_+}^{+\infty}\nu^{-2}\left(1+\frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-1/\xi-1}\mathrm{d}\nu\mathrm{d}r\mathrm{d}\xi \\
&= \frac{1}{m}\int_{-1/2}^{0}(1+2\xi)^{-1/2}\int_0^{+\infty} r^{3/2}e^{-r}\int_{-\xi(1+\xi)(x-u)}^{+\infty}\nu^{-2}\left(1+\frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-1/\xi-1}\mathrm{d}\nu\mathrm{d}r\mathrm{d}\xi \\
&+ \frac{1}{m}\int_{0}^{+\infty}(1+2\xi)^{-1/2}\int_0^{+\infty} r^{3/2}e^{-r}\int_{0}^{+\infty}\nu^{-2}\left(1+\frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-1/\xi-1}\mathrm{d}\nu\mathrm{d}r\mathrm{d}\xi \\
&= \frac{1}{m}\int_{-1/2}^{0}(1+2\xi)^{-1/2}\int_0^{+\infty} r^{3/2}e^{-r}\left[\frac{1}{(1+\xi)(x-u)}\left(1+\frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-1/\xi}\right]_{-\xi(x-u)\left(\frac{r}{m}\right)^{\xi}}^{+\infty}\mathrm{d}r\mathrm{d}\xi \\
&+ \frac{1}{m}\int_{0}^{+\infty}(1+2\xi)^{-1/2}\int_0^{+\infty} r^{3/2}e^{-r}\left[\frac{1}{(1+\xi)(x-u)}\left(1+\frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-1/\xi}\right]_{0}^{+\infty}\mathrm{d}r\mathrm{d}\xi \\
&= \frac{1}{m(x-u)}\int_{-1/2}^{+\infty}(1+\xi)^{-1}(1+2\xi)^{-1/2}\int_0^{+\infty} r^{3/2}e^{-r}\mathrm{d}r\mathrm{d}\xi \\
&= \frac{3\pi^{3/2}}{4m(x-u)} < \infty.
\end{aligned}
$$

Therefore, the posterior is proper for $n = 1$. It is well-known that it stays so for $n > 1$ as can be seen by induction. For instance for $n = 2$, the posterior writes

$$p(\boldsymbol{\theta} \mid x_1, x_2) \propto p(x_1, x_2 \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) = p(x_2 \mid \boldsymbol{\theta})p(x_1 \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(x_2 \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid x_1) \leq p(\boldsymbol{\theta} \mid x_1)$$

which is integrable.

Proof of Proposition 4 Similarly to the proof of Proposition 2, the aim is to show the existence of the following integral for any n:

$$C_n = \iiint_{\mathcal{S}} \frac{p_{\mathrm{PC}}(\xi \mid \lambda)}{\nu} e^{-r}\left(\frac{r}{m}\right)^n\left(\frac{\nu}{1+\xi}\right)^{-n}\prod_{i=1}^{n}\left(1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right)^{-1-\frac{1}{\xi}}\mathrm{d}r\mathrm{d}\nu\mathrm{d}\xi,$$

with $p_{\mathrm{PC}}(\xi \mid \lambda)$ defined in Equation (11), and $\mathcal{S}$ the following integration domain:

$$\mathcal{S} = \left\{(r, \nu, \xi) \in \mathbb{R}^3 \text{ s.t. } \xi < 1,\, r > 0,\, \nu \geq \{-\xi(1+\xi)((\max_i x_i) - u)\}_+\right\}.$$

In the general case for n we have

$$
\begin{aligned}
C_n &= \int_{-\infty}^{1}\int_{\{-\xi(1+\xi)(x-u)\}_+}^{+\infty}\frac{p_{\mathrm{PC}}(\xi \mid \lambda)}{\nu}\left(\frac{\nu}{1+\xi}\right)^{-n}\prod_{i=1}^{n}\left(1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right)^{-1-\frac{1}{\xi}}\int_0^{+\infty}\left(\frac{r}{m}\right)^n e^{-r}\mathrm{d}r\mathrm{d}\nu\mathrm{d}\xi \\
&= \frac{\Gamma(n+1)}{m^n}\int_{-\infty}^{1}\int_{\{-\xi(1+\xi)(x-u)\}_+}^{+\infty}\frac{p_{\mathrm{PC}}(\xi \mid \lambda)}{\nu}\left(\frac{\nu}{1+\xi}\right)^{-n}\prod_{i=1}^{n}\left(1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right)^{-1-\frac{1}{\xi}}\mathrm{d}\nu\mathrm{d}\xi \\
&= \frac{\Gamma(n+1)}{m^n}\int_{-\infty}^{1}\int_{\{-\xi(x-u)\}_+}^{+\infty}\frac{p_{\mathrm{PC}}(\xi \mid \lambda)}{\sigma}\sigma^{-n}\prod_{i=1}^{n}\left(1+\xi\left(\frac{x_i-u}{\sigma}\right)\right)^{-1-\frac{1}{\xi}}\mathrm{d}\sigma\mathrm{d}\xi.
\end{aligned}
$$

The remaining integral corresponds to the normalizing constant of the posterior distribution of a GPD model with a prior of the form $p(\sigma, \xi) \propto p(\xi)/\sigma$. Since $p(\xi)$ is a proper density, Theorem 1 in Northrop and Attalides (2016) concludes about the finiteness of the integral for any n. Note that this result remains true with $p_{\mathrm{PC}}(\xi \mid \lambda)$ replaced by a Laplace distribution as suggested in Section 3.2, as the prior on ξ remains proper.

Appendix C Additional experiments

C.1 Simulations using an Hamiltonian Monte Carlo algorithm

Hamiltonian Monte Carlo (HMC) (Neal, 2011) and its variants such as NUTS (Hoffman and Gelman, 2014) are MCMC methods with a Markov kernel based on trajectories of particles computed using Hamiltonian dynamics. Because of this, the performance of these methods is also sensitive to the choice to the parameterization (see Betancourt (2019) for a formalization of the problem). We performed the same experiments as those in Section 4.1 and Appendix C.2, using 500 NUTS iterations instead of 1 000 Metropolis–Hastings draws. The results obtained here are similar, and show that the orthogonal parameterization improves the efficiency of NUTS sampling. Figure C.7 illustrates the cases $\xi > 0$, with the same configuration as the one described in the first paragraph of Appendix C.2. We observe similar trends in this figure as those in Figure C.8: changing the value of m improves convergence, and using the orthogonal parameterization is even better. Moreover, NUTS seems to be more efficient on the three cases than with Metropolis–Hastings, as the chains seem to be less correlated compared to their equivalent in Figure C.8, and the ESS can even be greater than the number of draws.

C.2 Simulations on other maximum domains of attraction

We study the influence of parameterizations for MCMC convergence in cases where $\xi > 0$ and $\xi = 0$.

Example with $\xi > 0$ Here, we set $(m, u, \mu, \sigma, \xi) = (5, 10, 30, 15, 0.7)$, which leads to an expected number of observations of $r \approx 239$. Looking at autocorrelations, ESS and $\hat{R}(x)$ curves in Figure C.8, we can first confirm the result of Sharkey and Tawn (2017) about the inefficiency of Metropolis–Hastings on the original parameterization: high autocorrelations, high $\hat{R}(x)$ (around 1.7 for the highest) and almost zero ESS even after 1 000 iterations indicate a severe convergence issue. Changing the value of m before the MCMC algorithm as suggested by Sharkey and Tawn (2017) improves inference significantly. Still, considering the orthogonal parameterization is even more efficient, especially for the estimation of the tail parameter ξ: the autocorrelation reduces even more rapidly with the lag, and the ESS increases faster with the number of draws. With the recommendations of ESS ≥ 400 for estimation (Gelman et al., 2013), our experimental setup is satisfactory only in the orthogonal case because of ξ estimation. In contrast, more iterations are required to fulfil this condition for the parameterization recommended by Sharkey and Tawn (2017).

Example with $\xi = 0$ Finally when $\xi = 0$, the GPD and therefore the intensity $\Lambda(I_u)$ of the Poisson process defined in Section (1.1) reduces to an exponential model with location and scale parameters. Figure C.9 shows an example in this case with $(m, u, \mu, \sigma, \xi) = (20, 20, 25, 5, 0)$, leading to $r \approx 54$ expected observations. Similarly to the case $\xi < 0$ in Section 4.1, this example illustrates that updating m like Sharkey and Tawn (2017) is not beneficial for MCMC convergence. On the

other hand, one could surprisingly almost be satisfied with the original parameterization, despite ESS values for μ and σ that are still low after 1 000 iterations. The orthogonal parameterization, in the same way as in the two other maximum domains of attraction, is the most efficient one for the convergence of Metropolis–Hastings algorithm.

C.3 GPD and GEV case

In the particular case of GPD (defined in Equation (2)) that arises in the traditional peak over threashold model, the same observation can be made about the benefits of an orthogonal parameterization for (σ, ξ). More precisely, the transformation $(\nu, \xi) = (\sigma(1+\xi)), \xi)$ leads to an orthogonal Fisher information matrix for GPD (Chavez-Demoulin and Davison, 2005), and improves MCMC convergence as shown in Figure C.10. The same experimental setup as in the Poisson process case is used here, with a choice of $(\sigma, \xi) = (5, -0.1)$ and $u = 25$. Again, all plots in Figure C.10 show that the chains are satisfactory only in the case of an orthogonal parameterization, while the original parameterization requires more iterations to be effective for inference. Up to our knowledge, there is no orthogonal parameterization for the GEV likelihood known in the literature. However, it should be noted that the parameters of the Poisson process model (μ, σ, ξ) correspond to those of the block maxima framework with m blocks (see Section 1.1). Consequently, we should expect a similar convergence issue for parameters (μ, σ, ξ) with GEV likelihood, and therefore an improvement in the MCMC convergence with the use of the orthogonal parameterization (r, ν, ξ) of the Poisson model.

C.4 Replications and comparison with maximum likelihood

Despite the Bayesian paradigm comes with several benefits besides performance (briefly described in Section 1.1), one can be interested in the comparison with frequentist estimator like maximum likelihood estimation (MLE). From a frequentist point of view, this involves extracting a pointwise estimator from the posterior distribution, like the posterior mean, and replicate the experiment to estimate the mean squared error (MSE). The two steps of the Bayesian workflow we study here are expected to impact the performance of these estimators. A parameterization which leads to poor convergence of the MCMC chains will affect the accuracy of estimation, and the prior can add a bias that may or may not be advantageous to the estimation.

For different values of ξ_0 between -0.5 and 1, we replicate 100 times the following experiment (this range includes a large number of models and allows to have both Jeffreys and PC priors always defined): for $i = 1, \dots, 100$, we generate samples $\boldsymbol{x}_i$ according to a Poisson process distribution with parameters $(m, u, \sigma, \xi) = (1, 10, 15, \xi_0)$ and μ in a way such that the expected number of points is equal to $r = 100$:

$$\mu = u - \frac{\sigma}{\xi_0}(100^{-\xi_0} - 1).$$

Then, we run MCMC chains with the same configuration as in Section 4 and compute the posterior mean $\hat{\xi}_i = \mathbb{E}[\xi \mid \boldsymbol{x}_i]$. We these 100 experiments, we compute the MSE:

$$\mathrm{MSE}(\xi_0) = \frac{1}{100}\sum_{i=1}^{100}(\hat{\xi}_i - \xi_0)^2.$$

First, we compare the different parameterizations for the Poisson process with the same Jeffreys prior. Results are displayed in the left panel of Figure C.11, and illustrate the inaccuracy of the

frameworks without reparameterization and with the update of Sharkey and Tawn (2017), due to lack of convergence of MCMC. This issue is getting worse as ξ_0 increases, and a bias/variance decomposition of the MSE shows that it is mostly due to the variance term. Then, for the same orthogonal parameterization, we compare Jeffreys prior, PC prior with a choice of $\lambda = 10$, and the MLE for the Poisson process, implemented in the extRemes package (Gilleland and Katz, 2016). Results in the right panel of Figure C.11 show that the performance of the posterior mean estimation with Jeffreys prior is approximately the same as the MLE, except when ξ_0 is near $-1/2$ where the asymptote behaviour of Jeffreys favours the estimation. This shows that, despite the uninformative construction, this prior can favour a lot negative values of ξ_0 close to $-1/2$. The behaviour of PC prior is, as expected, penalizing the values of ξ far from $\xi_0 = 0$. When ξ_0 is around zero, this prior outperforms Jeffreys' one and MLE, but assuming a value near zero when $|\xi_0|$ is large can add a large bias.

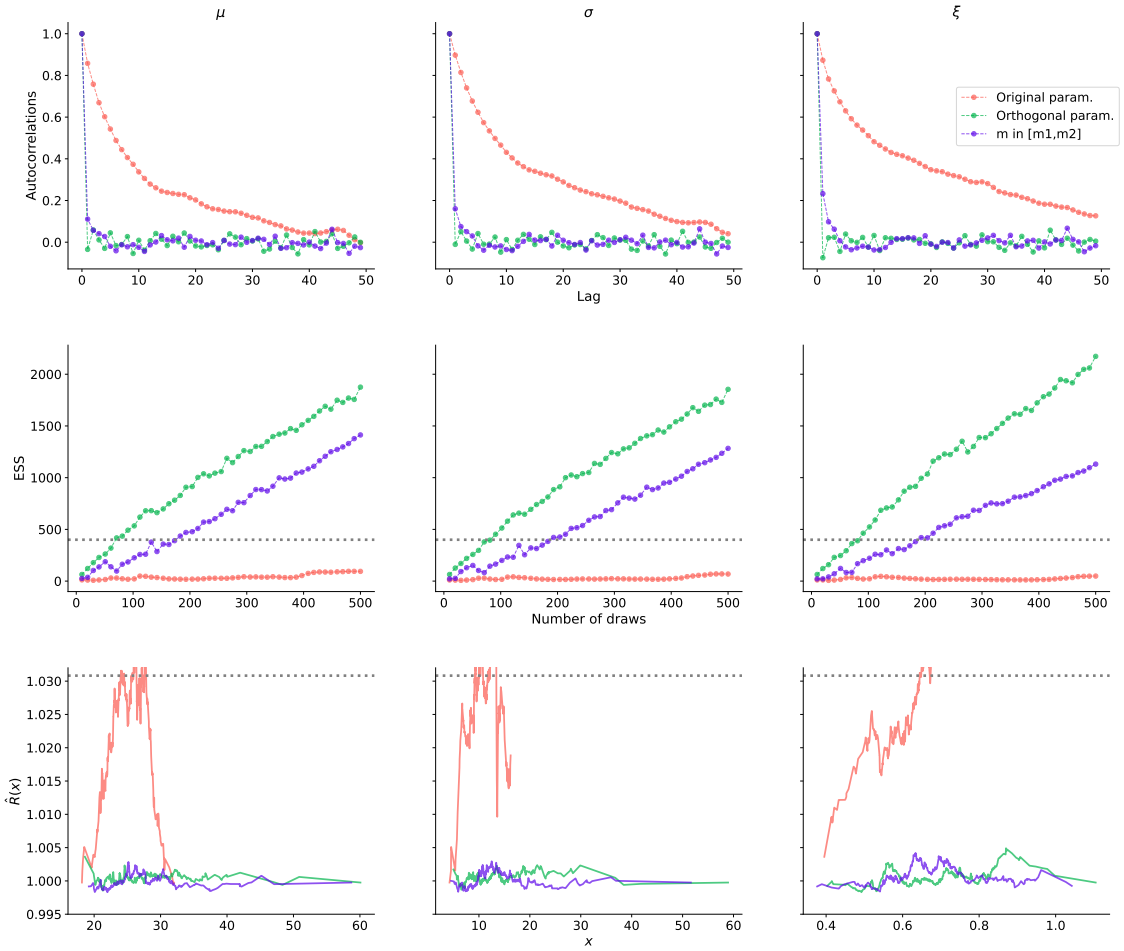


Figure C.7: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi > 0$, after 500 NUTS draws and a burn-in of 1 000, for three different parameterizations: the original one (in red), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in blue), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.031 (see Moins et al., 2022).

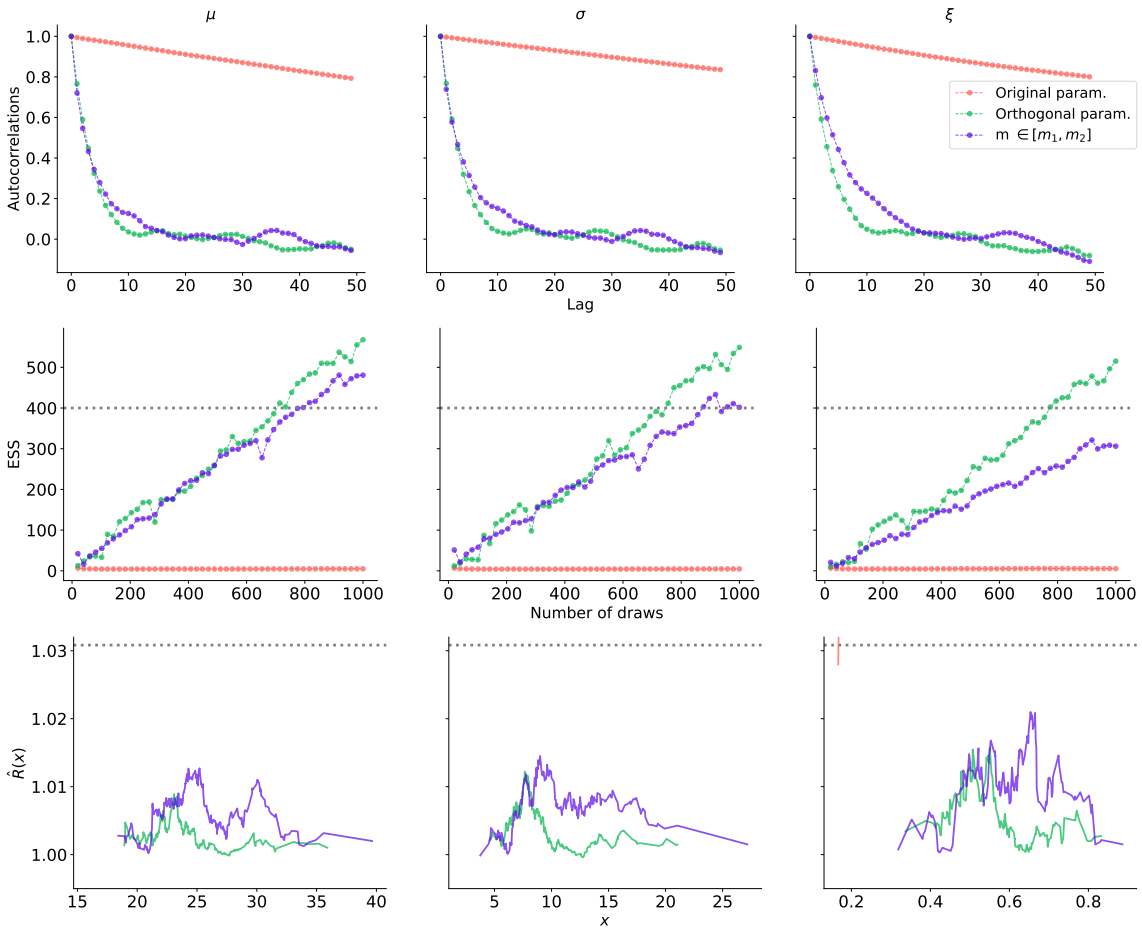


Figure C.8: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi > 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for three different parameterizations : the original one (in red), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in blue), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.031 (see Moins et al., 2022).

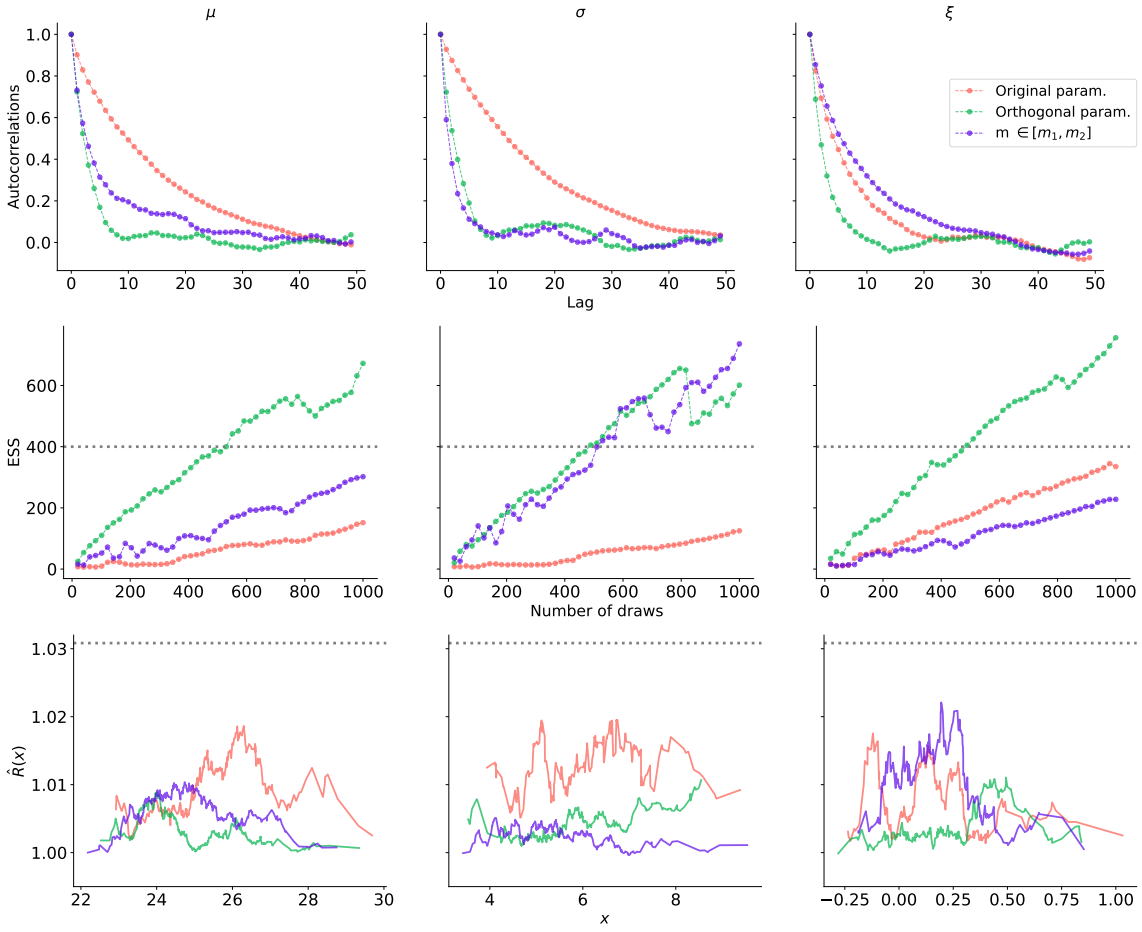


Figure C.9: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi = 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for three different parameterizations: the original one (in red), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in blue), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.031 (see Moins et al., 2022).

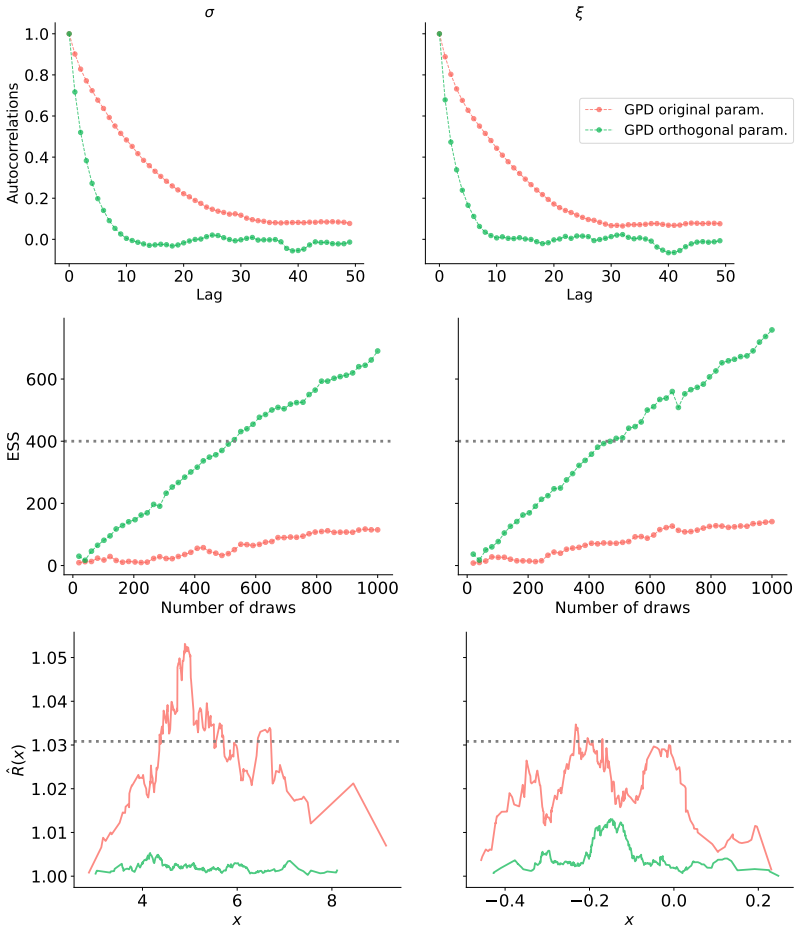


Figure C.10: Convergence diagnostic plots for GPD parameters (σ, ξ) with $\xi < 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for two parameterizations, the original (in red) and the orthogonal one (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.031 (see Moins et al., 2022).

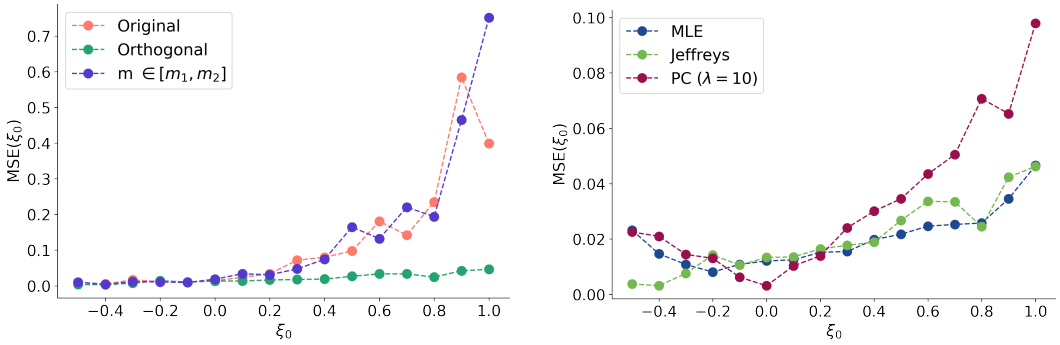


Figure C.11: Mean squared error (MSE) on the estimation of ξ for a true value $\xi_0 \in [-1/2, 1]$. The computation is done on 100 replications for each value of ξ_0. Left panel: different parameterizations under Jeffreys prior. Right panel different priors under orthogonal parameterization.