Extremes (2021) 24:325–348
https://doi.org/10.1007/s10687-020-00379-y

Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis

Axel Bücher[1] · **Jona Lilienthal**[2] · **Paul Kinsvater**[2] · **Roland Fried**[2]

Received: 16 May 2019 / Revised: 25 March 2020 / Accepted: 14 May 2020 /
Published online: 3 June 2020

Abstract

A common statistical problem in hydrology is the estimation of annual maximal river flow distributions and their quantiles, with the objective of evaluating flood protection systems. Typically, record lengths are short and estimators imprecise, so that it is advisable to exploit additional sources of information. However, there is often uncertainty about the adequacy of such information, and a strict decision on whether to use it is difficult. We propose penalized quasi-maximum likelihood estimators to overcome this dilemma, allowing one to push the model towards a reasonable direction defined a priori. We are particularly interested in regional settings, with river flow observations collected at multiple stations. To account for regional information, we introduce a penalization term inspired by the popular Index Flood assumption. Unlike in standard approaches, the degree of regionalization can be controlled gradually instead of deciding between a local or a regional estimator. Theoretical results on the consistency of the estimator are provided and extensive simulations are performed for the reason of comparison with other local and regional estimators. The proposed procedure yields very good results, both for homogeneous as well as for heterogeneous groups of sites. A case study consisting of sites in Saxony, Germany, illustrates the applicability to real data.

Keywords Regionalization · Index flood assumption · Generalized extreme value distribution · Consistency with rate · Tuning parameter selection

AMS 2000 Subject Classifications 62G32 · 62F12 · 62P12

✉ Jona Lilienthal
lilienthal@statistik.tu-dortmund.de

[1] Mathematisches Institut, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany

[2] Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany

1 Introduction

In flood frequency analysis, and more generally in statistics for extremes in hydrology (Katz et al. 2002), one is typically confronted with a (possibly non-stationary) version of the following problem: let $X_1, \ldots, X_n$ denote independent annual maximal river flows observed at a specific site and during the past n years, and let $F(x) = \mathbb{P}(X_i \leq x)$ denote their stationary cumulative distribution function (c.d.f.). The goal is to estimate a high quantile $q = F^{-1}(p)$, where typically the sample length n is small and the probability $p \in (0, 1)$ is high. This inconvenient imbalance results in estimators with a high variance and constitutes the main motivation for most of the statistical innovations in the field.

A widely accepted framework for the analysis of annual maxima, or more generally of block maxima, relies on the assumption that the c.d.f. F belongs to the 3-parametric generalized extreme value (GEV) distribution

$$G_\theta(x) = \exp\left[-\left(1 + \xi\,\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right] \text{ for } 1 + \xi\,\frac{x-\mu}{\sigma} > 0, \tag{1}$$

where the parameters $\theta = (\mu, \sigma, \xi)' \in \Theta = \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$ are called location, scale, and shape, respectively. The model is motivated by the fact that the members of the GEV family arise as the only possible limits in law of a block maximum $M_b = \max\{Z_1, \ldots, Z_b\}$ of independent (or weakly serially dependent) identically distributed random variables $Z_1, \ldots, Z_b$, after proper standardization and for block length $b \to \infty$ (de Haan and Ferreira (2006), Th. 1.1.3, and Leadbetter (1974), Th. 2.1). In much of the recent work related to climate change, the parameter vector θ is further assumed to depend on covariates, typically time and often in a parametric way (El Adlouni et al. 2007; Cannon 2009). See Serinaldi and Kilsby (2015) for a discussion on the merits and pitfalls of non-stationary models for extremes.

Being particularly interested in high quantiles (i.e., the right tail), note that the GEV family can handle a wide variety of right tail behaviour, with bounded right tails for $\xi < 0$, exponential tails for $\xi = 0$ and arbitrarily heavy tails for $\xi > 0$. The drawback of this flexibility shows up in the estimation of the parameter vector θ, particularly by a high estimation variance of the shape ξ resulting in a volatile quantile estimate. Different attempts have been made to reduce the estimation uncertainty for such estimation problems, in statistics for extremes in general and particularly in flood frequency analysis.

For instance, probability weighted moments or L-moments have been proposed as alternatives to moment or maximum likelihood (ML) estimators. Indeed, the former show a superior performance in typical small sample cases (Hosking et al. 1985), which has been mainly attributed to their restricted parameter space (Coles and Dixon 1999).

Alternative approaches are based on reducing the model complexity, for instance by restricting oneself to the two-parametric sub-family with a predefined shape like

$\xi = 0$, resulting in the location-scale Gumbel model (Lu and Stedinger 1992). The shortcoming of this approach is that only tails of one specific form (exponential if $\xi = 0$) are taken into account, which is not appropriate for many practical applications, in particular those that are primarily interested in the tails.

Finally, several attempts have been made to include additional sources of information into flood analyses. Throughout this paper, we will mainly address regional and seasonal methods, though other applications of the general methodology are possible. Regional methods require that observations from $d \geq 2$ river stations are available, with site-specific distributions denoted by F_j, $j = 1, \ldots, d$. The well-known Index Flood model (Dalrymple (1960), see also the monograph Hosking and Wallis (2005)) is then based on the assumption that the distribution at each station is the same except for some local scale. In other words, all local quantile functions are assumed to be identical to a regional quantile function H^{-1} except for the local scale $s_j = s(F_j) > 0$, that is

$$\mathcal{H}_{0,\mathrm{IF}} : \begin{cases} \exists \text{ c.d.f. } H \text{ and constants } s_j = s(F_j) > 0 \text{ such that} \\ F_j^{-1}(p) = s_j \cdot H^{-1}(p) \ \ \forall\, j = 1, \ldots, d, \ \ p \in [0, 1]. \end{cases} \tag{2}$$

Under this assumption, it is possible to reduce the variability of a quantile estimator at a specific site by taking observations from other sites into account (see Buishand (1991) for an application to precipitation extremes). Alternatively, seasonal methods do not only use time series on an annual scale but consider, say, monthly maximal flows, allowing for seasonal variability (Waylen and Mk 1982; Buishand and Demaré 1990; Baratti et al. 2012).

The two last-mentioned approaches, the reduction of local model complexity and the homogenization of a collection of stations, can be considered in the framework of *regularization*. Let F_n denote the empirical c.d.f. of the data $X_1, \ldots, X_n$ and suppose that one aims at minimizing some risk measure $R(\theta;\ F)$ with respect to a model parameter θ, where the c.d.f. F of the data is unknown. As for instance demonstrated in Vapnik (2000) by a simple regression example, minimizing the empirical counterpart $R(\theta;\ F_n)$ over the whole parameter space Θ is typically not the best strategy in finite samples. A more sophisticated and often preferable strategy (reducing possible overfitting) takes an additional penalty term $\Omega(\theta) \geq 0$ into account, which can be interpreted as measuring model complexity or representing a priori expert knowledge:

$$\hat{\theta}_\Omega = \operatorname*{arg\,min}_{\theta \in \Theta} R(\theta;\ F_n) + \Omega(\theta). \tag{3}$$

The idea of accounting for model complexity in the estimation of GEV parameters is not new. In fact, using the so-called cross-entropy risk $R(\theta;\ F) = -\mathbb{E}[\log g_\theta(X)]$, where g_θ is the density of G_θ from Eq. 1 and X is a random variable with $F(x) = \mathbb{P}(X \leq x)$, then minimizing the empirical cross-entropy $R(\theta;\ F_n) = -n^{-1} \sum_{i=1}^n \log g_\theta(X_i)$ with respect to θ is equivalent to ML estimation. When including a non-zero penalty, the resulting estimators are therefore called penalized maximum likelihood (PML) estimators. Coles and Dixon (1999) and Martins and Stedinger (2000) propose two slightly different estimators of GEV parameters of this

particular form Eq. 3, with a regularizer $\Omega(\theta)$ depending only on the shape ξ, thus aiming at ruling out unusual values of the shape parameter. However, no asymptotic theory is provided and it is unknown whether (and under what conditions) the estimators are consistent. The same is true for related approaches in extremes for hydrology, see, e.g., the PML estimators in Song et al. (2018) proposed for non-stationary Pearson-type 3 distributions. It is worthwhile to mention that, due to the fact that the support of the GEV distribution depends on the parameter, even the asymptotic behavior of the classical ML estimator is actually quite complicated, and has just recently been fully derived in Bücher and Segers (2017) and Dombry and Ferreira (2017).

The main contributions of this work are as follows: first of all, we present profound asymptotic results in a quite general multivariate setting, going far beyond the univariate settings mentioned in the previous paragraph. The main theoretical result is a consistency statement where the rate of convergence depends in an explicit way on the level of penalization. The results are partly similar to results in Pötscher and Leeb (2009) in the Gaussian case but the analysis is more difficult due to the non-smooth behaviour of the GEV distribution at the boundary of its support. Secondly, we illustrate the issue of choosing a suitable penalizing function Ω for some non-trivial problems with the prime example being flood frequency analysis based on the index flood assumption. Moreover, we propose a data-adaptive approach to select a tuning parameter that controls the level of penalization in finite samples. We illustrate that the proposed method performs very well compared to existing standard methods in an extensive simulation study, and that it yields easily interpretable results in a case study.

It is worthwhile to mention that the PML estimators considered in this paper may alternatively be interpreted (and even motivated) from a Bayesian perspective (for a related Bayesian approach to precipitation extremes see, e.g., Cooley et al. (2007)). In fact, under independence assumptions, a simple calculation shows that the PML estimator is actually equal to the posterior mode when assuming that θ is a random variable with prior density proportional to $\pi(\theta) = \exp(-\lambda\Omega(\theta))$. Hence, on the one hand, this paper partly offers an alternative view on Bayesian methods, and in particular provides a frequentist validation for them. On the other hand, the Bayesian perspective may also allow for an uncertainty assessment of the proposed procedure in terms of posterior distributions (see also Wood et al. (2017), and citations within). This paper being frequentist in nature, the latter approach is not pursued further here.

The remainder of this paper is organized as follows: Section 2 provides illustrations of possible applications of PML estimators in flood frequency analysis. Section 3 presents theoretical properties of such estimators in a general multivariate framework with GEV marginals. The degree of penalization is controlled by a hyperparameter, and the problem of its selection is treated in Section 4. An extensive simulation study in Section 5 compares the Index Flood penalization to estimators common in hydrology. A case study in Section 6 illustrates the applicability to hydrological data. Section 7 concludes this paper with a discussion of the most important findings. Proofs and additional simulation results can be found in an online supplement.

2 Regularization in flood frequency analysis

Within this section, we illustrate the broad applicability of PML techniques in flood frequency analysis. For illustrative purposes, we start with a simple approach based on penalizing unusual GEV shape parameters in a univariate setting. Then we discuss two possibilities to include additional data by jointly estimating the parameters at a set of stations using an Index Flood like penalization (adding regional information) and by using monthly instead of annual maxima (adding seasonal information).

2.1 Simple shape parameter penalization

Let $X_1, \ldots, X_n$ represent the data, consisting of independent and identically distributed observations with unknown distribution function $F(x) = \mathbb{P}(X_i \leq x)$. We are interested in the estimation of a high quantile $q = F^{-1}(p)$ from a rather small sample length n. Often enough, flood frequency analysts need to deal with $p \geq 0.99$ and $n \leq 50$.

Restriction to a 2-parametric sub-family of the GEV-model, like the Gumbel or a GEV distribution with a fixed shape parameter ξ_c, reduces the variance of a respective quantile estimator but possibly leads to a large bias. As a first application, we use penalization as an alternative to such a strict reduction of model complexity. More precisely, suppose that an expert claims that the true shape parameter ξ_0 is close to $\xi_c = 0.2$. This knowledge may be reflected by choosing a penalty term of the form $\Omega_\lambda(\theta) = \lambda(\xi - \xi_c)^2$ with hyperparameter $\lambda \geq 0$ reflecting our confidence in this prior belief, and by considering the PML estimator

$$\hat{\theta}_\lambda \in \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log g_\theta(X_i) - \lambda\,(\xi - \xi_c)^2. \tag{4}$$

If the expert was perfectly sure that actually $\xi_0 = \xi_c$ holds, we should choose $\lambda = \infty$ and thus enforce an estimate of θ with third component $\hat{\xi} = \xi_c$ (using the convention that $\infty \cdot 0 = 0$). Alternatively, we can select any value $0 \leq \lambda < \infty$ reflecting the uncertainty in the expert's prior information with $\lambda = 0$ leading to the ordinary ML estimator.

For further insight, we present the outcome of a small simulation experiment. Figure 1 depicts common empirical performance measures of estimators $\hat{q}_\lambda = G^{-1}_{\hat{\theta}_\lambda}(0.99)$ with $\hat{\theta}_\lambda$ from Eq. 4, $\xi_c = 0.2$, and increasing values of λ. The measures are computed from 10 000 independent samples of size $n = 50$ each with true parameter $\theta_0 = (\mu_0, \sigma_0, \xi_0)' = (2, 1, 0.4)'$. Note that our prior information reflected by $\Omega_\lambda(\theta) = \lambda(\xi - 0.2)^2$ is not centered around the true value of $\xi_0 = 0.4$. The (almost) unbiasedness of the ML estimator (for $\lambda = 0$) is outweighed by larger variability. Increasing the value of λ can be interpreted as trading variance for bias. In this example, the estimator $\hat{q}_\lambda$ with $\lambda = 20$ performs best in terms of empirical mean squared error, and every value $\lambda > 0$ leads to better performance than $\lambda = 0$, although $\xi_0 = 0.4$ is not close to our a priori guess $\xi_c = 0.2$. Also note that neither $\lambda = 0$ nor $\lambda = \infty$ are optimal in this scenario. This can be explained by the strong imbalance between a small sample length and a comparably high quantile.

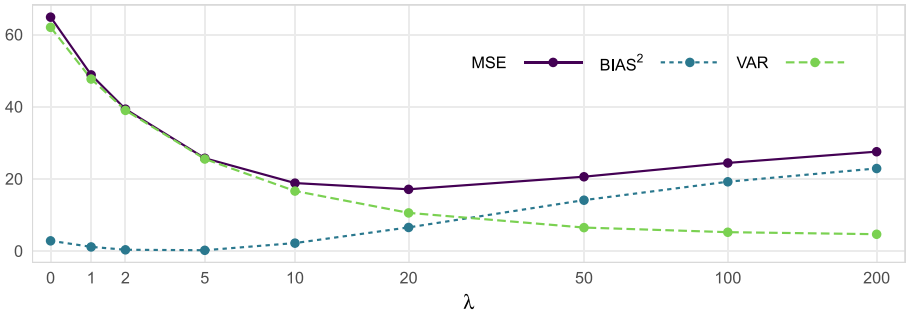

Fig. 1 Empirical MSE, squared bias, and variance of a 0.99 quantile estimate from PML parameter estimation with penalized deviations of the shape parameter to a ξ_c. Even though $\xi_c = 0.2$ does not correspond to the true $\xi_0 = 0.4$, regularization is still beneficial ($n = 50$)

A more comprehensive simulation study reveals that the previous findings strongly depend on the sample length, the true parameters, the object of interest and the expert guess/the penalty. In particular, in situations where ξ_c is much larger than ξ_0, the ML estimator ($\lambda = 0$) may still be the best estimator regarding MSE, with values of λ close to zero leading to acceptable estimation as well. Selecting a suitable value of λ is the most critical task in application of the PML method and will be discussed in Section 4 below.

2.2 Penalization inspired by the index flood model

Practitioners typically have data from several stations in the same region available. The local record lengths often vary substantially over the stations, typically with different start times (times of gauge installations) and a common end time. The data scheme can hence be written as

$$\left.\begin{array}{rl} X_{1,1},\ X_{2,1},\ X_{3,1},\ X_{4,1},\ X_{5,1},\ \ldots,\ X_{n,1} & \sim F_1 \\ X_{a_2+1,2},\ X_{a_2+2,2},\ X_{a_2+3,2},\ \ldots,\ X_{n,2} & \sim F_2 \\ \ddots & \vdots \\ X_{a_d+1,d},\ X_{a_d+2,d},\ \ldots,\ X_{n,d} & \sim F_d \end{array}\right\} \begin{array}{l} \text{observations} \\ \text{from } d \text{ sites} \end{array} \tag{5}$$

where $X_{i,j}$ shall represent the annual maximum observation at station j in year i of the observation period. Moreover, n denotes the longest record length, $a_j + 1$ ($0 \le a_j \le n$) the individual start times and $n_j = n - a_j$ the individual record lengths. For ease of presentation, we arranged the samples in Eq. 5 such that the first station corresponds to that with the full sample length of $n_1 = n$, i.e. $a_1 = 0$.

We assume that the random vectors $X_i = (X_{i,1}, \ldots, X_{i,d})'$, consisting of possibly partially observed values for the different years, are independent and identically distributed with GEV margins $F_j \in \{G_{\theta_j} : \theta_j = (\mu_j, \sigma_j, \xi_j)' \in \Theta\}$ for all $j = 1, \ldots, d$. Note that we neither assume the d components $X_{i,j}$ for the same time point i to be independent nor that we impose a specific model for the spatial dependence.

Recall the Index Flood assumption from Eq. 2. If we additionally assume that the common distribution is a member of the GEV family, i.e., $H = G_{\theta_0}$ for certain parameters $\theta_0 = (\mu_0, \sigma_0, \xi_0)' \in \Theta$, the hypothesis $\mathcal{H}_{0,\mathrm{IF}}$ is equivalent to $\theta_j = (\mu_j, \sigma_j, \xi_j)'$ satisfying

$$\frac{\mu_1}{\sigma_1} = \ldots = \frac{\mu_d}{\sigma_d} = \delta_0 \text{ and } \xi_1 = \ldots = \xi_d = \xi_0 \text{ for some } \delta_0, \xi_0 \in \mathbb{R}. \tag{6}$$

A straightforward combination of the Index Flood principle and penalization techniques suggests to penalize deviations between $\delta_j = \mu_j/\sigma_j$ and δ_0 and deviations between ξ_j and ξ_0. Because δ_0 and ξ_0 are not known, we replace them by approximations δ_c and ξ_c, which can be chosen as weighted means, $\delta_c = \sum_{j=1}^{d} w_j \delta_j$ and $\xi_c = \sum_{j=1}^{d} w_j \xi_j$ with weights $w_j = n_j / \sum_{j'=1}^{d} n_{j'}$, or using a priori knowledge. A suitable penalization is given by

$$\Omega_\lambda(\theta) = ((\delta_1 - \delta_c)^2, \ldots, (\delta_d - \delta_c)^2, (\xi_1 - \xi_c)^2, \ldots, (\xi_d - \xi_c)^2)\lambda, \tag{7}$$

with hyperparameter $\lambda = (\lambda_{11}, \ldots, \lambda_{1d}, \lambda_{21}, \ldots, \lambda_{2d})' \in [0, \infty]^{2d}$. This results in the penalized quasi ML estimator (simply denoted by PML throughout)

$$\hat{\theta}_\lambda \in \underset{\theta \in \Theta^d}{\arg\max} \sum_{j=1}^{d} \sum_{i=a_j+1}^{n} \log g_{\theta_j}(X_{i,j}) - \sum_{j=1}^{d} \left\{ \lambda_{1j} \, (\delta_j - \delta_c)^2 + \lambda_{2j} \, (\xi_j - \xi_c)^2 \right\}. \tag{8}$$

The term quasi refers to the fact that the likelihood is derived under the additional assumption of spatially independent observations which is actually not necessary for consistency of the estimator, see Section 3.

In this application, increasing the hyperparameters λ reflects stronger belief in $\xi_j \approx \xi_c$ and $\delta_j \approx \delta_c$ for all $j = 1, \ldots, d$ or weaker certainty about the quality of the local estimator. In fact, both options of regular flood frequency analysis, calculation of local or regional estimates, are included as special cases when choosing $\lambda = 0$ or $\lambda \to \infty$, respectively. The elegance of this approach lies in the fact that strange local estimates are effectively ruled out without relying completely on the restrictive application of the Index Flood model or an arbitrary initial guess. The performance of this estimator in finite samples will be analysed in detail by simulations in Section 5, and by a real-data application in Section 6.

2.3 Penalization inspired by seasonal smoothness assumptions

An analysis that considers seasonal or monthly maxima instead of annual maxima allows to expand the available information and can improve the estimation of very high quantiles. The underlying motivation is that, due to different flood origins like snow melt or heavy rainfall, stochastic characteristics (smoothly) vary over the course of a year. Fischer et al. (2016) analysed such a seasonal modeling and found generalized extreme value distributions appropriate to describe the distribution of seasonal maxima. In this section we expand on this by penalizing differences in the shape parameter of monthly maximal flows assuming a GEV distribution.

At a particular station, the observed monthly maximal flows are denoted by $M_1^{(m)}, \ldots, M_n^{(m)} \sim F^{(m)}$, $m = 1, \ldots, 12$. Under the assumption of independence of the monthly maxima, quantiles of the annual maximal flows $X_i = \max\left\{M_i^{(1)}, \ldots, M_i^{(12)}\right\}$ are given by

$$F^{-1}(p) = \left(F^{(1)} \cdot \ldots \cdot F^{(12)}\right)^{-1}(p). \tag{9}$$

For illustrative purposes, we consider the distribution of the monthly maxima $F^{(m)}$ to be given by GEV distributions G_{θ_m}, $m = 1, \ldots, 12$, despite the fact that the GEV assumption is not necessarily met on such a fine scale. The vector of unknown model parameters $\theta = \left(\theta_1', \ldots, \theta_{12}'\right)'$ is estimated by

$$\hat{\theta}_\lambda \in \underset{\theta \in \Theta^{12}}{\arg\max} \sum_{m=1}^{12} \sum_{i=1}^{n} \log g_{\theta_m}\left(M_i^{(m)}\right) - \Omega_\lambda(\theta), \tag{10}$$

using a penalty Ω_λ that prefers gradually changing shape parameters $\xi_1, \ldots, \xi_{12}$ over the year. More specifically, we set

$$\Omega_\lambda(\theta) = \Omega_\lambda\left(\xi_1, \ldots, \xi_{12}\right) = \lambda\Big\{\sum_{m=1}^{11}(\xi_m - \xi_{m+1})^2 + (\xi_{12} - \xi_1)^2\Big\}, \tag{11}$$

which implies a natural periodicity of one year. Note that we could have also incorporated similar penalties for location and scale parameters.

Figure 2 shows the outcome of a simulation experiment based on 10 000 independent samples of $n = 50$ independent GEV observations per month with $\mu_0 = 2$, $\sigma_0 = 1$ and shapes following a sine curve $\xi_0^{(m)} = 0.35 + 0.25\sin(m\pi/6 + 3)$, $m = 1, \ldots, 12$, with a period of one year. The boxplots illustrate the distribution of the

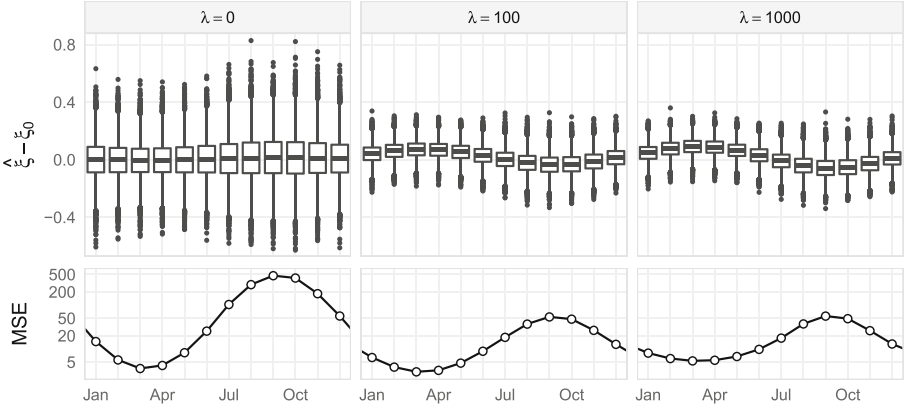

Fig. 2 Top: Boxplots of the difference between shape parameter estimate and true shape parameter for every month and three choices of λ. Bottom: Empirical MSE of each month's 0.99 quantile estimate. The choice $\lambda = 0$ leads to the smallest overall bias of the shape parameter but $\lambda = 100$ to the smallest MSE of the 0.99 quantile estimate

difference between the shape estimate and the true shape parameter for each month and different penalties $\lambda \in \{0, 100, 1000\}$. The empirical MSE of the respective 0.99 quantile estimate of each month are depicted below. The regular ML estimator ($\lambda = 0$) leads to the lowest overall bias, but trading some variance for bias, a much smaller MSE can be achieved by $\lambda = 100$. This choice also leads to the smallest MSE of the yearly 0.99 quantile estimate calculated using Eq. 9 among the considered penalties (MSE of 226 compared to 1728 for $\lambda = 0$ and 232 for $\lambda = 1000$). An approach using only yearly maxima would have resulted in an MSE of 839, so the seasonal model yields a substantial gain in this situation.

2.4 Further extensions

The examples presented before assumed stationary distributions (over the years), but, due to known or unknown causes like river regulations or climate change, the assumption of stationarity is often not justified. PML estimators can be applied in such scenarios. An intuitive way to model a time-dependent distribution $F_t = G(\theta_t)$ is by splitting the time span $\{1, \ldots, n\}$ into b blocks for which we assume stationarity, i.e.,

$$\theta_t = \kappa(t) = \begin{cases} \theta_1 = (\mu_1, \sigma_1, \xi_1)', & t \in [i_0, i_1), \\ \vdots & \vdots \\ \theta_b = (\mu_b, \sigma_b, \xi_b)', & t \in [i_{b-1}, i_b], \end{cases} \tag{12}$$

for given $1 = i_0 < i_1 < \cdots < i_{b-1} < i_b = n$ (based on our simulation experience with the stationary setting, we would recommend to at least choose block lengths $i_j - i_{j-1} \geq 20$, but this recommendation should be taken with some care). It is reasonable to penalize differences between parameters of consecutive blocks, for example $\Omega_\lambda(\theta) = \lambda \sum_{j=1}^{b} (\xi_j - \xi_{j-1})^2$, possibly in addition to other penalizations. Since the main focus of this paper is to analyse PML estimators in the context of regionalization, we restrict to stationarity in the following sections.

In the previous three subsections, we have also focused on squared distances in the penalization term. As an alternative, one could use absolute differences as in LASSO regression (Tibshirani (1996)), which lead to a built-in variable selection in regression problems by automatically setting coefficients to zero. In the seasonal context illustrated in Section 2.3 using absolute distances would result in an estimator similar to the Fused LASSO (Tibshirani et al. 2005). In our applications, however, there is no particular advantage in setting individual parameters exactly equal to other parameters or to pre-described values. Throughout our simulation study described in Section 5, we have checked the performance of absolute differences (similar to a LASSO approach) and of a combination of absolute and squared differences (similar to the so-called elastic net) in different settings, but these choices lead to inferior empirical MSEs as compared to quadratic differences and also to higher computation times. We concentrate on quadratic differences in this work and use the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS), a quasi-Newton method, for the optimization of the objective function Eqs. 4, 8 or 10, respectively, to be given in a general form in Eq. 13 in the next section.

3 Theoretical results

We show that the PML estimator exists (i.e., the maximization problem has a solution) and is consistent under fairly general conditions on the penalty. We also provide a result about the rate of consistency, which turns out to depend explicitly on the strength of penalization. All proofs are deferred to Section A in the supplementary material.

Let $X_1, \ldots, X_n$ with $X_i = (X_{i1}, \ldots, X_{id})'$ be an i.i.d. sequence from $X = (X_1, \ldots, X_d)'$, a d-dimensional random vector with marginal cumulative distribution functions denoted by $F_1, \ldots, F_d$. We assume that the marginal laws are from the GEV-family, that is, there exists $\theta_{0j} = (\mu_{0j}, \sigma_{0j}, \xi_{0j})' \in \Theta_{-1} = \mathbb{R} \times (0, \infty) \times (-1, \infty)$ such that $F_j = G_{\theta_{0j}}$, for $j = 1, \ldots, d$. Note that the parameter set Θ_{-1} is restricted to $\xi_{0j} > -1$ since otherwise the classical ML estimator for the GEV parameters is not consistent (Dombry 2015). The dependence between the coordinates of X is left unspecified.

Note that the setting of Section 2.2 fits into this framework, with d denoting the number of sites, as long as $a_j = 0$ for $j = 2, \ldots, d$ (the results can however be easily extended to situations with $n' = n - a_d \to \infty$). The setting of Section 2.3 is accomplished with $d = 12$; additionally, the coordinates of X are assumed to be stochastically independent then.

Let $\theta_0 = (\theta_{01}', \ldots, \theta_{0d}')' \in \Theta_{-1}^d$ denote the stacked vector of true marginal parameters. A generic vector of marginal parameters will be denoted by $\theta = (\theta_1', \ldots, \theta_d')'$, with $\theta_j = (\mu_j, \sigma_j, \xi_j)'$. Let $\hat{\theta}$ denote any local maximum of the function

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \log g_{\theta_j}(X_{ij}) - \frac{1}{n} \lambda_n' \Omega(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(X_i) - \frac{1}{n} \lambda_n' \Omega(\theta), \quad (13)$$

where $\Omega : \Theta_{-1}^d \to [0, \infty)^m$ denotes an arbitrary penalty function.

The following first main result shows that there always exists a strongly consistent local maximizer, as soon as the smoothing parameter is of smaller order than n. Similar results have been obtained for Lasso-type estimators in a linear regression model in Knight and Fu (2000), although their results are easier to obtain due to the convexity of their criterion function. Our proof is based on similar arguments as in Dombry (2015).

Proposition 1 (Strong Consistency) *Let K denote an arbitrarily large compact subset of Θ_{-1}^d, containing θ_0 in its interior. Suppose that the penalty Ω is continuous. Then, provided $\lambda = \lambda_n$ satisfies $\|\lambda_n\| = o(n)$ as $n \to \infty$, any estimator $\hat{\theta}_n$ such that*

$$Q_n(\hat{\theta}_n) = \sup_{\theta \in K} Q_n(\theta), \quad (14)$$

such maximizers always existing, is strongly consistent for θ_0, as $n \to \infty$.

While the estimator is strongly consistent for any smoothing parameter of the order $o(n)$, it turns out that the rate of convergence of $\|\hat{\theta}_n - \theta_0\|$ to zero in fact depends on the precise order of the smoothing parameter. The following second main results

shows that we obtain the usual parametric rate for $\|\lambda_n\| = O(\sqrt{n})$, and smaller rates for $\|\lambda_n\|$ between $n^{1/2}$ and n, asymptotically. Similar results have been obtained for Lasso-type estimators in simple linear regression models in Pötscher and Leeb (2009), Section 4. For technical reasons, we restrict ourselves to the reduced parameter set $\Theta_{-1/2} = \mathbb{R} \times (0, \infty) \times (-1/2, \infty)$, as this is the set where the GEV family is differentiable in quadratic mean and the usual ML estimator is $\sqrt{n}$-consistent and asymptotically normal, see Bücher and Segers (2017).

Proposition 2 (Rate of Convergence) *Suppose that the conditions of Proposition 1 are met, with K denoting a compact subset of $\Theta^d_{-1/2}$ containing θ_0 in its interior. Additionally, let Q be Lipschitz-continuous on K. Then, as $n \to \infty$,*

$$\|\hat{\theta}_n - \theta_0\| = \begin{cases} O_{\mathbb{P}}(n^{-1/2}) & \text{if } \|\lambda_n\| = O(\sqrt{n}), \\ O_{\mathbb{P}}(n^{-1/2+\kappa}) & \text{if } \|\lambda_n\| = O(n^{1/2+\kappa}) \text{ for some } \kappa \in [0, 1/2). \end{cases} \tag{15}$$

Regarding the proof, the regime $\|\lambda_n\| = o(\sqrt{n})$ may be treated with Theorem 5.52 and Corollary 5.53 in Van der Vaart (2000), see also Bücher and Segers (2017), Proposition D.1. For $\|\lambda_n\|$ of larger order, a suitable adaptation of Theorem 5.52 in Van der Vaart (2000) is needed, which may be interesting in its own right; this is Proposition 6 in the supplementary material. An empirical illustration of the consistency statements with rate can be found in Section B in the supplementary material.

A next desirable result would consist of deriving the precise asymptotic distribution of $\hat{\theta}_n$. This, however, is beyond the scope of this paper and left for future research; note that the problem is difficult due to the fact that the support of the GEV-family depends on the parameter (whence standard theory does not apply).

4 Hyperparameter selection

In this section, strategies to select appropriate values of λ are discussed. We restrict attention to estimator Eq. 8 inspired by the index flood model, but similar approaches are applicable to the seasonal smoothing estimator Eq. 10 or the general estimator Eq. 14.

We propose a cross-validation procedure based on the empirical cross-entropy. The set of observed years, $I = \{1, \ldots, n\}$, is partitioned evenly into K subsets, $I_1, \ldots, I_K \subset I$, that do not necessarily consist of consecutive years and are chosen randomly. Let $F_n^{(k)}$ be the empirical c.d.f. of the k-th subset and let $\hat{\theta}_\lambda^{(-k)} = ((\hat{\theta}_{\lambda,1}^{(-k)})', \ldots, (\hat{\theta}_{\lambda,d}^{(-k)})')'$ be the estimator of θ_0 calculated without the data of the k-th group. Select the parameter $\lambda \in [0, \infty]^m$ that minimizes the sum of empirical cross-entropies $R(\hat{\theta}_\lambda^{(-k)}, F_n^{(k)})$ over all groups, i.e.,

$$\begin{aligned} \lambda^{CV} &= \operatorname*{arg\,min}_{\lambda \in [0,\infty]^m} \sum_{k=1}^{K} R(\hat{\theta}_\lambda^{(-k)}, F_n^{(k)}) \\ &= \operatorname*{arg\,max}_{\lambda \in [0,\infty]^m} \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{\{j : a_j < i\}} \log g_{\hat{\theta}_{\lambda,j}^{(-k)}}(X_{i,j}). \end{aligned} \tag{16}$$

Throughout our simulation experiments and applications, we choose the often-recommended $K = 10$ groups (Hastie et al. (2009), page 242). The much higher computational cost of a Leave-one-out cross-validation using $K = n$ did not lead to a better quality of the selected hyperparameter in our experiments.

If λ is high dimensional, the optimization of Eq. 16 can become very complex or even not feasible. In this case, constraints on λ can simplify calculations. More precisely, for some $m' \leq m$, let $\tau : [0, \infty]^{m'} \to [0, \infty]^m$ be a given fixed function. The resulting constrained estimator associated with τ is written as $\lambda^{CV} = \tau(\lambda^{CV}_{cons})$ with

$$\lambda^{CV}_{cons} = \underset{\lambda \in [0,\infty]^{m'}}{\arg\max} \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{\{j: a_j < i\}} g_{\hat{\theta}^{(-k)}_{\tau(\lambda),j}}(X_{i,j}). \tag{17}$$

The most simple constraint is equality of all hyperparameters, i.e., $\lambda_{1j} = \lambda_{2j} = \lambda$ for all $j = 1, \ldots, d$, which is achieved using $\tau(\lambda) = (\lambda, \ldots, \lambda)'$, $\lambda \in [0, \infty]$. We refer to hyperparameters derived using this τ as λ^{CV}_{global}.

Note that equality of all hyperparameters does not imply that the penalization effect is the same for sites with different record lengths. Indeed, the log-likelihood part of Eq. 8 consists of different numbers of observations while the penalization term is independent of the observation length. Hence, the ratio between those two parts is different according to the length of records, penalizing sites with few records (relatively) more than sites with many records.

Alternatively, to have stronger differences in the penalization effect but still a feasible dimension of λ, the constraint $\lambda_{1j} = \lambda_{2j} = \lambda_j$ for all $j = 1, \ldots, d$ can be used, and is achieved by $\tau(\lambda_1, \ldots, \lambda_d) = (\lambda_1, \ldots, \lambda_d, \lambda_1, \ldots, \lambda_d)'$. We denote this selection as λ^{CV}_{local}.

As we will see in the results of the simulation study, globally selected hyperparameters tend to have high bias and low variance while individually selected hyperparameters tend to have low bias and high variance. To investigate whether combinations of the local and global λ result in a better estimation, we also consider $\lambda^{CV}_{comb,\alpha} = \alpha\lambda^{CV}_{local} + (1 - \alpha)\lambda^{CV}_{global}$, $\alpha \in [0, 1]$.

In regional flood frequency analysis, groups of stations are often built based on site characteristics like mean elevation, mean slope, or catchment area. An alternative to purely observation-based cross-validation procedures could be to map a measure of the goodness-of-fit of a given site to a given group to the λ-space $[0, \infty]^m$. We will briefly investigate this approach in the case study in Section 6.

5 Simulation study

In this section we compare the performance of the PML estimator for regional flood quantile estimation with standard methods in this field.

5.1 Scenarios

We generate several synthetic data sets of different types and different lengths. Heterogeneity can manifest in many different forms and is hard to capture systematically.

To include a wide variety of heterogeneity structures, we consider four different types: (I) a setting in which the sites are divided into two groups (called "groups"), (II) sites with linearly varying parameters ("linear"), (III) a setting in which four sites vary in different directions from the remaining, equally distributed sites ("single"), and (IV) a setting with parameters that are arranged in a spherical fashion ("spherical"). All sites follow GEV(μ_j, σ_j, ξ_j) distributions with the location parameter of station $j = 1, \ldots, d$ set to $\mu_j = 5j$. The location-scale ratio $\delta_j = \mu_j/\sigma_j$ (and hence the scale parameter) and the shape parameter ξ_j of station j are selected using the following formulas in the four settings (I)-(IV):

$$\delta_j = 1.8 + r \times \Delta_1(j, d), \qquad \xi_j = 0.2 + 2r \times \Delta_2(j, d) \tag{18}$$

$$\text{with} \begin{cases} \Delta_1(j,d) = \Delta_2(j,d) = \text{sign}(\frac{j-1}{d-1} - \frac{1}{2}), & \text{(I)} \\ \Delta_1(j,d) = \Delta_2(j,d) = \frac{j-1}{d-1} - \frac{1}{2}, & \text{(II)} \\ \Delta_1(j,d) = \mathbb{1}_{\{1,2\}}(j) - \mathbb{1}_{\{3,4\}}(j),\ \Delta_2(j,d) = \mathbb{1}_{\{1,3\}}(j) - \mathbb{1}_{\{2,4\}}(j), & \text{(III)} \\ \Delta_1(j,d) = \cos(\frac{j}{d}2\pi),\ \Delta_2(j,d) = \sin(\frac{j}{d}2\pi) & \text{(IV)} \end{cases}$$

and with parameter $r \in \mathbb{R}_+$ controlling the degree of heterogeneity, $\mathbb{1}_A$ denoting the indicator function of a set A and sign the signum function. Figure 3 illustrates the four settings, for the choices of $r = 0.1$ (I), $r = 0.2$ (II) and $r = 0.15$ (III and IV). The central coordinate $(1.8, 0.2)$ was chosen because it is an average coordinate in the case study presented in Section 6. We select record lengths between 20 and 100 observations and $d = 12$ stations. Quantile estimates of different heights are calculated from $B = 5000$ replications of each scenario using the methods described in the following section.

For the ease of a clear presentation, we only present results in spatially independent settings. Alternative simulation scenarios based on dependent data (with dependency described by a Gumbel copula) did not exhibit any fundamental qualitative differences, aside from increased estimator variances for all methods.

5.2 Methods

We compare local and regional methods that are based either on ML estimation (including our proposed penalized estimator) or L-moments.

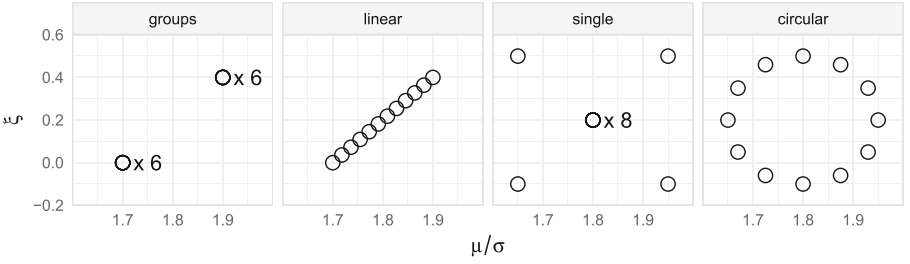

Fig. 3 Representation of the four data settings. Sites differ group-wise, linearly, in a circular fashion, or with single outliers in terms of loc-scale-ratio and shape parameter

L-moment based estimators are very common in hydrology, see Hosking (1990) for an introduction. The local L-moment method, denoted by `l-local`, calculates L-moments for each site individually and converts them to GEV parameters $\hat{\theta}_j^L = (\hat{\mu}_j^L, \hat{\sigma}_j^L, \hat{\xi}_j^L)'$, $j = 1, \ldots, d$. The regional L-moment method, `l-regional`, uses the well-known regional flood frequency approach of Hosking and Wallis (2005), which is based on the Index Flood model given in Eq. 2. L-moments are calculated from the normalized series X_{ij}/s_j, $i = 1, \ldots, n_j$, with individual Index Floods s_j, $j = 1, \ldots, d$, being calculated as local arithmetic means. Regional L-moments are built as weighted means of these, with weights equal to the record lengths. Regional GEV parameters $\hat{\theta}_R = (\hat{\mu}_R, \hat{\sigma}_R, \hat{\xi}_R)'$ are calculated by converting the regional L-moments to GEV parameters. Local parameter estimates are then given through $\hat{\theta}_j^{LReg} = (\hat{\mu}_R\, s_j, \hat{\sigma}_R\, s_j, \hat{\xi}_R)'$, $j = 1, \ldots, d$. Note that Hosking and Wallis (2005) describe a much more comprehensive procedure, beginning with data screenings, identifications of homogeneous regions, and tests to check assumptions. We only concentrate on the data information pooling scheme in this study.

The local ML approach (denoted as `ml-local`) calculates ML estimates at each site individually by optimizing

$$\hat{\theta}_j^{ML} = \operatorname*{arg\,max}_{\theta \in \Theta} \sum_{i=a_j+1}^{n} \log g_\theta(X_{i,j}), \quad j = 1, \ldots, d. \tag{19}$$

Starting values for the numerical optimization are chosen from L-moments.

Our proposed method is the PML estimator described in Eq. 8. Throughout the optimization, we fix δ_c and ξ_c using weighted means of local L-estimates $\delta_c = n^{-1} \sum_{j=1}^{d} n_j \frac{\hat{\mu}_j^L}{\hat{\sigma}_j^L}$ and $\xi_c = n^{-1} \sum_{j=1}^{d} n_j \hat{\xi}_j^L$. This reduces the optimization problem to an individual maximization at each site:

$$\begin{aligned} \hat{\theta}_\lambda &= \operatorname*{arg\,max}_{\theta \in \Theta^d} \sum_{j=1}^{d} \sum_{i=a_j+1}^{n} \log g_{\theta_j}(X_{i,j}) - \sum_{j=1}^{d} \left(\lambda_{1j}\, (\delta_j - \delta_c)^2 + \lambda_{2j}\, (\xi_j - \xi_c)^2 \right) \\ &= \begin{pmatrix} \operatorname*{arg\,max}_{\theta_1 \in \Theta} \sum_{i=a_1+1}^{n} \log g_{\theta_1}(X_{i,1}) - \lambda_{11}\, (\delta_1 - \delta_c)^2 - \lambda_{21}\, (\xi_1 - \xi_c)^2 \\ \vdots \\ \operatorname*{arg\,max}_{\theta_d \in \Theta} \sum_{i=a_d+1}^{n} \log g_{\theta_d}(X_{i,d}) - \lambda_{1d}\, (\delta_d - \delta_c)^2 - \lambda_{2d}\, (\xi_d - \xi_c)^2 \end{pmatrix}. \end{aligned} \tag{20}$$

To determine appropriate hyperparameters λ we use cross-validation as described in Section 4 with $K = 10$ subsets. We use and compare the constrained hyperparameters λ_{global}^{CV}, λ_{local}^{CV}, as well as combinations $\lambda_{comb,\alpha}^{CV}$ with $\alpha \in \{0.25, 0.5, 0.75\}$. The methods will be denoted by `pml-gl`, `pml-ll`, `pml-cl-0.25`, `pml-cl-0.5`, or `pml-cl-0.75`, respectively.

All parameter estimates $\hat{\theta}$ are converted to quantile estimates by $\hat{q} = F_{\hat{\theta}}^{-1}(p)$, $p \in (0, 1)$.

5.3 Performance measures

To assess the quality of the methods we use common performance measures. Let $q_j = q_j(F_j)$ be a specific quantile of a distribution F_j and $\hat{q}_{b,j} = \hat{q}_{b,j}(\hat{\theta}_{\lambda,b})$ the corresponding estimation in sample $b = 1, \ldots, B$. For each method we calculate the average empirical relative mean squared error as

$$\text{relMSE} = d^{-1} \sum_{j=1}^{d} B^{-1} \sum_{b=1}^{B} \frac{(\hat{q}_{b,j} - q_j)^2}{q_j^2}. \tag{21}$$

We also examine the composition of this measure by calculating the mean empirical relative squared bias and mean empirical relative variance as

$$\text{relSqBias} = d^{-1} \sum_{j=1}^{d} \left(B^{-1} \sum_{b=1}^{B} \frac{\hat{q}_{b,j} - q_j}{q_j} \right)^2, \tag{22}$$

$$\text{relVar} = d^{-1} \sum_{j=1}^{d} B^{-1} \sum_{b=1}^{B} \left(\frac{\hat{q}_{b,j} - B^{-1} \sum_{b'=1}^{B} \hat{q}_{b',j}}{q_j} \right)^2. \tag{23}$$

5.4 Results

Figure 4 displays the relative MSE of the 0.99 quantile estimation for the PML methods with different hyperparameters in the linear and the single setting. The two settings not displayed are qualitatively comparable to the linear one. The global λ-selection, which selects the same hyperparameter for all sites, is the best choice in most of these situations. The relative MSE tends to get worse if a higher proportion of the local selection is used, with the only exception being the single setting with a high degree of heterogeneity. In this case, the locally chosen hyperparameters typically differ a lot so that improvements over equally chosen hyperparameters are possible. Since the improvement is not large however, we stick with λ_{global} for PML estimation in the following.

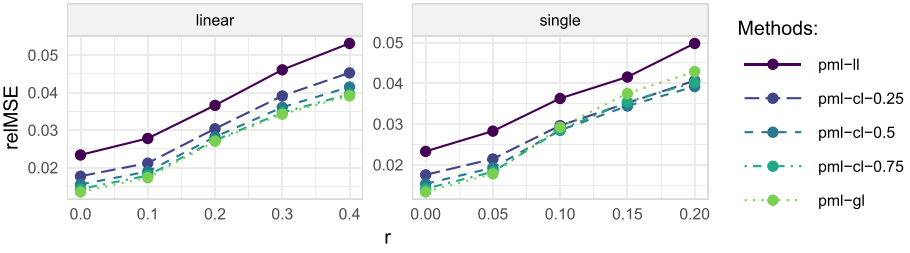

Fig. 4 Relative MSE depending on the heterogeneity r for different λ-selections ($n = 80$). The two settings not displayed are qualitatively comparable to the linear case

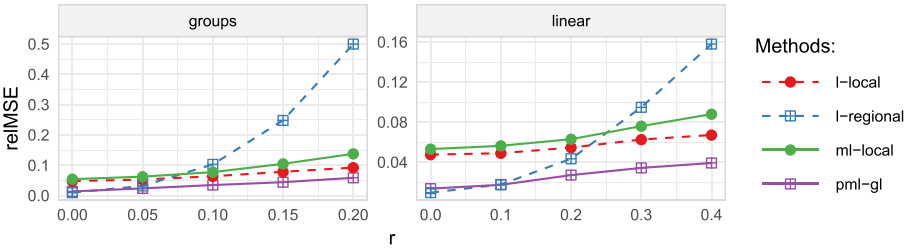

Fig. 5 Relative MSE of the 0.99 quantile estimators depending on the distance r for two settings and records of length $n = 80$

Figure 5 depicts the relative MSE of the estimates for the 0.99 quantile for record lengths of $n = 80$ and two settings. These illustrations are representative also for other quantiles, record lengths (as we will see later), and the other two settings. Both L-moment based methods perform well for their intended application, the regional one for homogeneous groups (small r) and the local one for heterogeneous groups (large r), but they lack quality if they are applied to the contrary situation. The PML estimator overcomes this problem by allowing to gradually choose between local or regional estimation. Using the globally selected hyperparameter λ it performs best or close to the best in all these situations, independently of the degree of heterogeneity r. The local L-moment based estimation outperforms the local ML based one in all settings considered here. As already discussed in Hosking et al. (1985), this is likely due to the short record length.

The top panels of Fig. 6 show the influence of the record length on the relative MSE in the linear setting for three degrees of heterogeneity. The local ML method fails for record lengths smaller than, say, $n = 40$ but it catches up with increasing

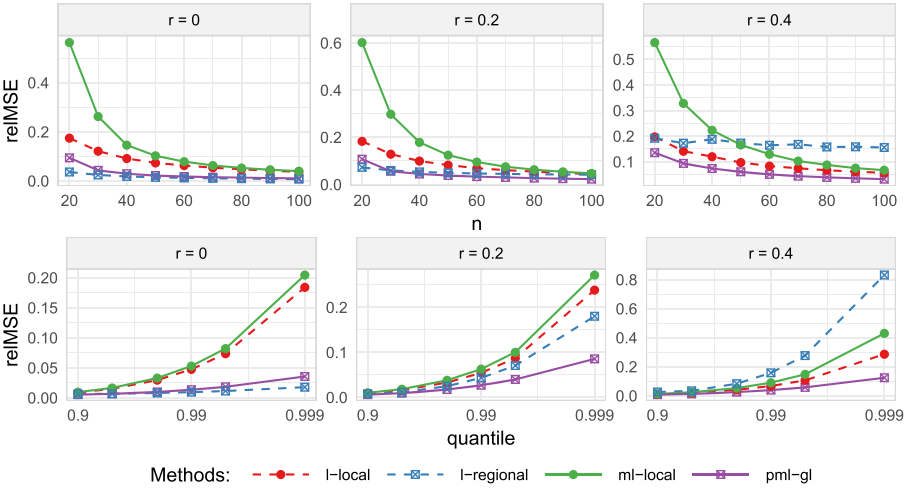

Fig. 6 Results for the linear setting. Top: Relative MSEs for the 0.99 quantile and different record lengths n. Bottom: Relative MSE as a function of the estimated quantile

record length, while the L-moment estimations are not that much influenced by small record lengths. The PML estimator gives good results for record lengths larger than $n = 30$ and is nearly as good as the regional L-moment estimator in homogeneous groups ($r = 0$) and surpasses all other methods in groups of higher heterogeneity.

The bottom panels of Fig. 6 show the MSE of the estimation of different quantiles in the linear setting for $n = 80$. The methods show stable relative performances for all quantiles and each heterogeneity. For homogeneous groups, the local methods show much larger MSE than the regional ones. As opposed to the regional L-moment estimator, the PML estimator remains the best choice among these methods as the heterogeneity increases.

Figure 7 finally splits the MSE into the squared bias and the variance. The squared bias increases rapidly with increasing heterogeneity for the regional L-moment method, while for the other methods it is rather small as compared to the variance. The variance is substantially smaller for the regional estimators than for the local ones, with a small advantage for the regional L-moment estimator in this respect.

Overall, the PML estimator combines a small squared bias with a low variance, which results in a good relative MSE. The proposed cross-validation procedure is able to provide hyperparameters that adapt to local or regional solutions depending on the data situation and can reduce the relative mean square error substantially in this way.

6 Case study

We illustrate the application of our PML estimator with a case study. The data set consists of flood peaks (maximal water discharge in m^3/sec) at 26 stations in the Elbe river

basin in Saxony, Germany, located at the north side of the Ore Mountains (with a mountaintop of 1244 m a.s.l.) and its foothills. The sites differ in mean elevation (from 168 m to 754 m a.s.l.) and catchment area (from around 36 km^2 to 5433 km^2) and consist of record lengths between 64 and 103 years.

We begin by illustrating the seasonal estimation method from Section 2.3 using monthly flood peaks at Rothenthal, a rather small catchment located in the Ore Mountains on the border between Germany and Czech Republic. The data set consists of

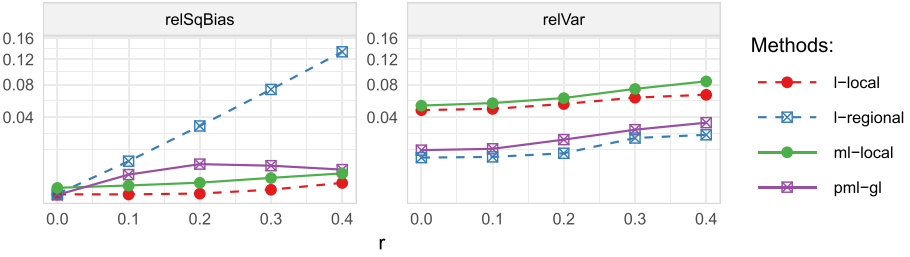

Fig. 7 Squared Bias and Variance of 0.99 quantile estimators depending on the heterogeneity r in the linear setting, $n = 80$

measurements for 84 years. The top panel of Fig. 8 contains the peaks separated by month. The estimator defined in Eq. 10 was calculated for different values of λ. The resulting shape parameter estimates are given in the bottom panel of Fig. 8. The regular ML estimate varies sharply and seems to be strongly influenced by single events; e.g. the shape of August is significantly higher than the shape of July although the distributions of the peaks look similar except for one very high event in August. Penalization leads to less extreme estimates and to a much smoother estimation curve. A 10-fold cross-validation was calculated using formula Eq. 16 and found $\lambda = 44.72$ to be the best choice (filled points in the bottom panel of Fig. 8). The cross-validated solution has a clear seasonal variability but avoids spikes or extreme estimates.

Next, we focus on the PML estimator in a regional setting based on annual maxima, as described in Section 2.2. Section 5 illustrates that the PML estimator for regional estimation yields comparably good results both in homogeneous and moderately heterogeneous situations, which is why small to moderate deviations from the homogeneity assumption can be tolerated when using this estimator. In order to protect against heavy deviations from homogeneity, it may however be advantageous to perform a group building process first. For that purpose, site characteristics (catchment area, mean elevation, proportion of forest area, stream density, length of stream network) are used to construct two groups by an application of k-means clustering on standardized site characteristics. One resulting group (mostly) contains sites with small catchment areas located at higher elevations of the Ore Mountains, while the other group includes sites with bigger catchment areas further downstream. Smaller

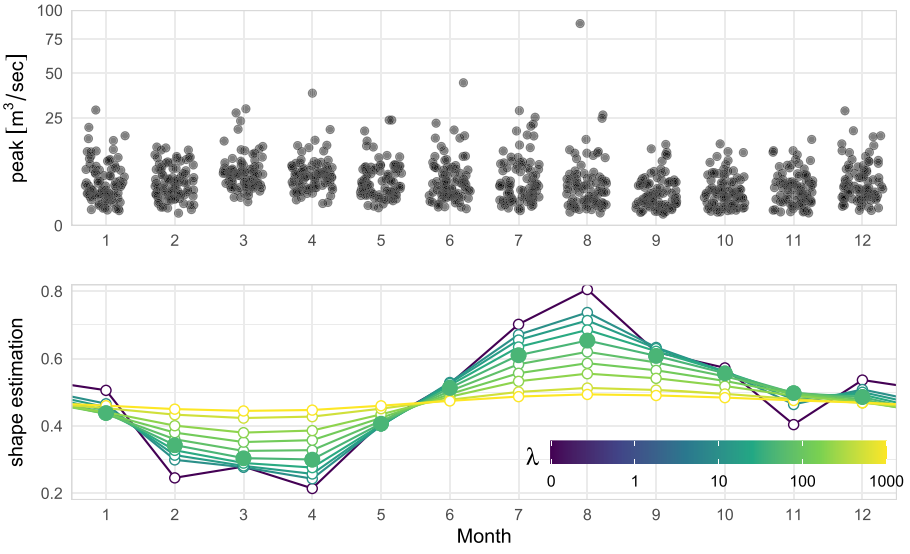

Fig. 8 Top: Monthly maximal discharges for each calendar month. Bottom: Shape parameter estimates using a seasonal penalization term. Filled points indicate the estimates based on the cross-validated hyperparameter

catchments are more strongly affected by single events and therefore often feature larger shape parameters in a GEV model, so that the grouping appears to be reasonable. To analyse the influence of the group-building process, the estimates are calculated once with a single group containing all sites and once after division into these two groups. Subsequently, d denotes the number of sites of the respective group under consideration; thus, its meaning may change from line to line.

The PML estimator of Eq. 20 is calculated for each group (or for all sites together) with a globally cross-validated λ (i.e. $\lambda_{1j} = \lambda_{2j} = \lambda \; \forall j = 1, \ldots, d$). Regarding the choice of δ_c and ξ_c, preliminary simulation results showed that selecting δ_c and ξ_c as weighted means of the corresponding local values results in rather ragged estimation paths $\lambda \mapsto (\hat{\gamma}_j(\lambda), \hat{\mu}_j/\hat{\sigma}_j(\lambda))'$. Much smoother paths are obtained by fixing the group centres δ_c and ξ_c at pre-specified weighted means of local L-moment estimates throughout the optimization. We therefore choose to present results for the latter approach only. The selected hyperparameters from the cross-validation are $\lambda^{CV}_{global} = 0.79$ if no grouping is applied and $\lambda^{CV}_{1,global} = 0.25$ and $\lambda^{CV}_{2,global} = 1.02$ if the sites are grouped. Figure 9 shows the respective estimates for both cases. In both plots, the lines indicate all estimates obtained by the PML estimator using $\lambda \in [0, \infty)$, with the local ML estimate (i.e. $\lambda = 0$) being the most outward point of the line. The bold points indicate the estimates chosen by the cross-validated λ^{CV}_{global}. Without grouping, the estimates vary moderately around the centre, clearly less than ordinary ML estimates would do. With two groups, there are clear differences: the first group (filled circles) has a medium level of regionalization, resulting in estimates in the middle of the path. Regionalization is much stronger for the other group, with all estimates being closer to the centre of the group.

Finally, we want to give a small example of how additional information can be used to improve the hyperparameter cross-validation. For that purpose we use a constraint function τ as in Section 4 in which we incorporate information about the dissimilarity of the sites to their group. The respective calculations are done for both groups separately. To measure the dissimilarities, we calculate the Euclidean distances $dist_j$, $j = 1, \ldots, d$, of each site to the mean of the corresponding

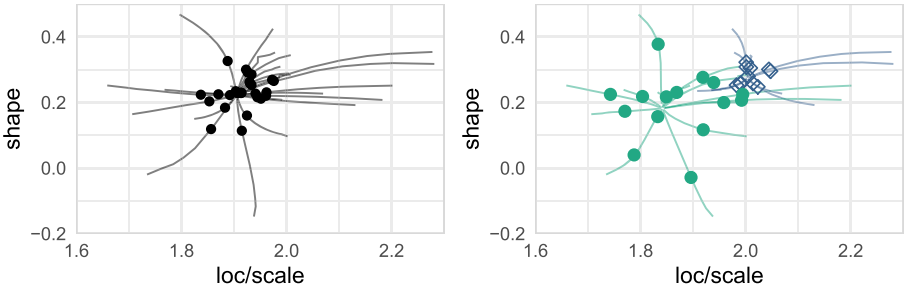

Fig. 9 PML estimates without (left) and with group-building (right). Lines mark all possible estimates for different hyperparameters, points indicate the cross-validated solution

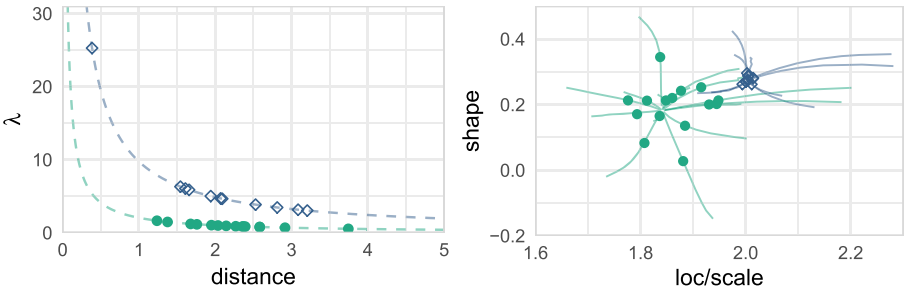

Fig. 10 Left: Mappings $dist \mapsto \lambda_\ell^{cons}/dist$ for $\lambda_1^{cons} = 1.03$ (green line, Group 1) and $\lambda_2^{cons} = 8.46$ (blue line, Group 2). The dots and diamonds indicate the individual sites in the respective groups. Right: Estimates using the hyperparameters $\lambda_\ell = \tau(\lambda_\ell^{cons})$

group in the space of the standardized site characteristics that were used for the k-means group building process. The constraint function τ is now constructed with two aspects in mind: first, we want to ensure that the final hyperparameter $\lambda = (\lambda_{11}, \lambda_{21}, \ldots, \lambda_{d1}, \lambda_{d2})' = \tau(\lambda^{cons})$ allows for an individual degree of regionalization $\lambda_{1j} = \lambda_{2j} = \lambda_j$ at each site and, second, that these λ_j have a reciprocal relationship to the dissimilarity $dist_j$. Hence, we set, for $\lambda^{cons} \in [0, \infty]$,

$$\begin{aligned}\tau(\lambda^{cons}) &= \tau_{(dist_1,\ldots,dist_d)}(\lambda^{cons}) \\ &= (\lambda^{cons}/dist_1, \ldots, \lambda^{cons}/dist_d, \lambda^{cons}/dist_1, \ldots, \lambda^{cons}/dist_d)'. \quad (24)\end{aligned}$$

A suitable value of λ^{cons} is found by applying formula Eq. 17 and the final hyperparameter is then selected as $\lambda = \tau(\lambda^{cons})$.

The obtained cross-validated hyperparameters are given by $\lambda_1^{cons} = 1.03$ and $\lambda_2^{cons} = 8.46$ for Group 1 and 2, respectively. In the left panel of Fig. 10, we depict the mappings $dist \mapsto \lambda_\ell^{cons}/dist$ for $\ell = 1, 2$, with the dots and diamonds representing the sites in Group 1 and 2, respectively. It can be seen that the final hyperparameters $\lambda_j = \lambda_{1j} = \lambda_{2j}$ are comparable for Group 1, while the variation is larger within Group 2, with one outlying site. In the right panel of Fig. 10 the corresponding estimates are given. Since Group 1 is mapped to small hyperparameters, the estimates are further away from the group centre. Group 2 is mapped to higher hyperparameters and has estimates close to the centre. These findings are similar to the previous ones, but with an even more regionalized second group.

Finally, Fig. 11 presents 95%-confidence intervals of site-specific 0.99 quantile estimates that are calculated applying a non-parametric resampling technique. More precisely, we create bootstrap samples by randomly drawing n years of the original dataset with replacement, calculate the estimates using the different methods in each bootstrap sample, and use the empirical 0.025 and 0.975 quantiles as confidence interval limits. For comparison, confidence intervals based on the local L-moment estimator, the ML estimator and the regional L-moment estimator using the same groups as our PML estimator are added as well (the latter is calculated using the dissimilarity information $dist_j$ as described above). The lengths of the confidence intervals based on PML estimation are shorter than those of local estimations and comparable to the size of regional L-moments. This indicates that the variability of

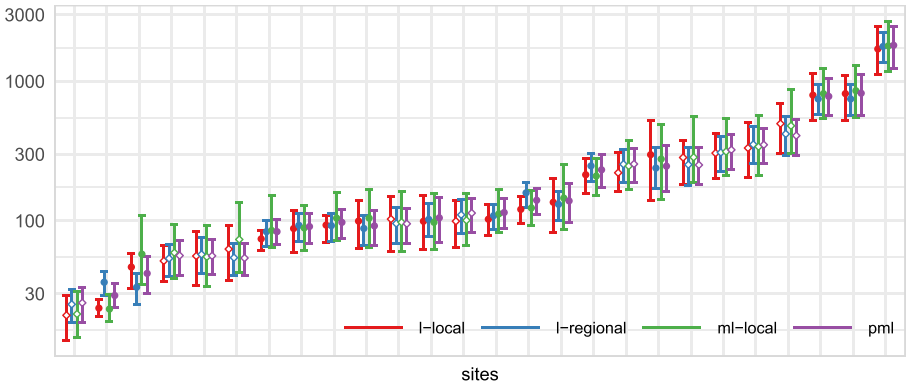

Fig. 11 Bootstrap confidence intervals of 0.99 quantile estimates using different estimators at all 26 sites. The intervals around filled dots and diamonds belong to the sites of Group 1 and 2, respectively

the PML estimator is similar to a regional procedure while maintaining individual estimations.

7 Discussion

This paper discusses PML estimators in extreme value models and provides theoretical large sample results for a rather general GEV framework. We prove strong consistency if the hyperparameter is of order $o(n)$ and show how the rate of convergence depends on the order of the hyperparameter.

Applications cover simple constraints on the shape parameter, seasonal constraints, and an Index Flood like regularization for regional flood frequency analysis. The latter is of particular interest and analysed using synthetic data in a simulation study and real data in a case study. The penalization term is chosen to represent the well-known index flood model by penalizing deviations from local parameter estimates to regionally calculated ones. A hyperparameter controls the influence of that term and thus the balance between local and regional estimates. In contrast to former methods, this enables us to adjust the degree of regionalization.

A crucial point in regularization techniques is the choice of hyperparameters, with common approaches being based on cross-validation procedures. Through simulations we have found that in our short record scenarios a globally selected hyperparameter (i.e. the same parameter for each site) is usually advantageous over selecting an individual parameter. The only setting in which this was not the case is a scenario in which the majority of sites is completely homogeneous and only few outlying sites differ from them. In this case the optimal hyperparameters differ a lot so that improvements over equally chosen hyperparameters are possible.

The main result of the simulation study is that the PML estimator generally provides competitive or even better quantile estimates as compared to other methods when there is uncertainty about the homogeneity of the group of sites. While the local L moment estimator and the regional L moment estimator using the Index Flood

model offer good results in each of the situation they are designed for, they lack quality if the situation is not clear or misspecified. The PML estimator overcomes this problem by allowing to gradually choose between local or regional estimation.

The real-world applicability has been demonstrated at a set of 26 gauges in Germany which were divided into two groups based on site-characteristics. Using this example we have shown how surrogate information like the distance of the stations to the center of the group in the space of site characteristics can be used to derive hyperparameters. The latter provides a promising alternative to observation-based cross-validation in situations of short record lengths.

The paper leaves several opportunities for further research. On the theoretical side, the asymptotic distribution of the PML estimator could be investigated. On the methodological side, extensions to the peaks-over-threshold approach might be of interest. In terms of applications, the approach described in Section 2.4 to cope with possible non-stationarities deserves a comprehensive investigation. Regarding regional flood frequency analysis, further investigations could concern the possibility that each site is not only penalised to the centre of one group but to multiple groups. Indeed, it seems more realistic that, for each site, there is no native membership to one group but different degrees of membership to several groups.

Acknowledgements Open Access funding provided by Projekt DEAL. The authors are grateful to the associate editor and two anonymous referees, whose comments on an earlier version of this manuscript lead to a significant improvement. Furthermore, the authors would like to thank Professor Andreas Schumann and his research group from the Department of Civil Engineering, Ruhr-University Bochum, Germany, for supplying the data used in the Case Study. Financial support of the Deutsche Forschungsgemeinschaft (SFB 823) is gratefully acknowledged.

References

Baratti, E., Montanari, A., Castellarin, A., Salinas, J., Viglione, A., Bezzi, A.: Estimating the flood frequency distribution at seasonal and annual time scales. Hydrol. Earth Syst. Sci. **16**(12), 4651–4660 (2012). https://doi.org/10.5194/hess-16-4651-2012

Bücher, A., Segers, J.: On the maximum likelihood estimator for the generalized extreme-value distribution. Extremes **20**(4), 839–872 (2017). https://doi.org/10.1007/s10687-017-0292-6

Buishand, T.A.: Extreme rainfall estimation by combining data from several sites. Hydrolog. Sc. J. **36**(4), 345–365 (1991). https://doi.org/10.1080/02626669109492519

Buishand, T.A., Demaré, G.R.: Estimation of the annual maximum distribution from samples of maxima in separate seasons. Stoch. Hydrol. Hydraul. **4**(2), 89–103 (1990). https://doi.org/10.1007/BF01543284

Cannon, A.J.: A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. Hydrol. Process. **24**(6), 673–685 (2009). https://doi.org/10.1002/hyp.7506

Coles, S.G., Dixon, M.J.: Likelihood-based inference for extreme value models. Extremes **2**(1), 5–23 (1999). https://doi.org/10.1023/A:1009905222644

Cooley, D., Nychka, D., Naveau, P.: Bayesian spatial modeling of extreme precipitation return levels. J. Amer. Stat. Assoc. **102**(479), 824–840 (2007). https://doi.org/10.1198/016214506000000780

Dalrymple, T.: Flood-frequency analyses manual of hydrology: Part 3. Tech. rep., USGPO (1960)

de Haan, L., Ferreira, A.: Extreme value theory: An introduction. Springer, Berlin (2006)

Dombry, C.: Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. Bernoulli **21**(1), 420–436 (2015). https://doi.org/10.3150/13-BEJ573

Dombry, C., Ferreira, A.: Maximum likelihood estimators based on the block maxima method. arXiv:1705.00465 (2017)

El Adlouni, S., Ouarda, T.B.M.J., Zhang, X., Roy, R., Bobée, B.: Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. Water Resour. Res. **43**(3), W03410 (2007). https://doi.org/10.1029/2005WR004545

Fischer, S., Schumann, A., Schulte, M.: Characterisation of seasonal flood types according to timescales in mixed probability distributions. J. Hydrology **539**, 38–56 (2016). https://doi.org/10.1016/j.jhydrol.2016.05.005

Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media, Berlin (2009)

Hosking, J.R.M.: L-moments: Analysis and estimation of distributions using linear combinations of order statistics. J. Royal Stat. Soc. B **52**(1), 105–124 (1990). https://doi.org/10.1111/j.2517-6161.1990.tb01775.x

Hosking, J.R.M., Wallis, J.R.: Regional frequency analysis: An approach based on L-moments. Cambridge University Press, Cambridge (2005)

Hosking, J.R.M., Wallis, J.R., Wood, E.F.: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics **27**(3), 251–261 (1985). http://www.jstor.org/stable/1269706

Katz, R.W., Parlange, M.B., Naveau, P.: Statistics of extremes in hydrology. Adv. Water Resour. **25**(8), 1287–1304 (2002). https://doi.org/10.1016/S0309-1708(02)00056-8

Knight, K., Fu, W.: Asymptotics for lasso-type estimators. Ann. Stat. **28**(5), 1356–1378 (2000). https://doi.org/10.1214/aos/1015957397

Leadbetter, M.R.: On extreme values in stationary sequences. Z. Wahrscheinlichkeitstheor Verw. Geb. **28**, 289–303 (1974). https://doi.org/10.1007/BF00532947

Lu, L.H., Stedinger, J.R.: Variance of 2-parameter and 3-parameter gev pwm quantile estimators - formulas, confidence-intervals, and a comparison. J. Hydrol. **138**(1-2), 247–267 (1992). https://doi.org/10.1016/0022-1694(92)90167-T

Martins, E.S., Stedinger, J.R.: Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. Water Resour. Res. **36**(3), 737–744 (2000). https://doi.org/10.1029/1999WR900330

Pötscher, B.M., Leeb, H.: On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. J. Multivar. Anal. **100**(9), 2065–2082 (2009). https://doi.org/10.1016/j.jmva.2009.06.010

Serinaldi, F., Kilsby, C.G.: Stationarity is undead: Uncertainty dominates the distribution of extremes. Adv. Water Resour. **77**, 17–36 (2015). https://doi.org/10.1016/j.advwatres.2014.12.013

Song, X., Lu, F., Wang, H., Xiao, W., Zhu, K.: Penalized maximum likelihood estimators for the nonstationary pearson type 3 distribution. J. Hydrol. **567**, 579–589 (2018). https://doi.org/10.1016/j.jhydrol.2018.10.035

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. B **58**(1), 267–288 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. Royal Stat. Soc. Series B (Stat. Methodol.) **67**(1), 91–108 (2005). https://doi.org/10.1111/j.1467-9868.2005.00490.x

Van der Vaart, A.W.: Asymptotic statistics, vol. 3. Cambridge University Press, Cambridge (2000)

Vapnik, V.N.: The nature of statistical learning. Springer, New York (2000)

Waylen, P., Mk, W.: Prediction of annual floods generated by mixed processes. Water Resour. Res. **18**(4), 1283–1286 (1982). https://doi.org/10.1029/WR018i004p01283

Wood, S.N., Li, Z., Shaddick, G., Augustin, N.H.: Generalized additive models for gigadata: modeling the uk black smoke network daily data. J. Am. Stat. Assoc. **112**(519), 1199–1210 (2017). https://doi.org/10.1080/01621459.2016.1195744