Extremes manuscript No.
(will be inserted by the editor)

A weighted composite log-likelihood approach to parametric estimation of the extreme quantiles of a distribution

Michael L. Stein

Received: date / Accepted: date

Abstract Extreme value theory motivates estimating extreme upper quantiles of a distribution by selecting some threshold, discarding those observations below the threshold and fitting a generalized Pareto distribution to exceedances above the threshold via maximum likelihood. This sharp cutoff between observations that are used in the parameter estimation and those that are not is at odds with statistical practice for analogous problems such as nonparametric density estimation, in which observations are typically smoothly downweighted as they become more distant from the value at which the density is being estimated. By exploiting the fact that the order statistics of independent and identically distributed observations form a Markov chain, this work shows how one can obtain a natural weighted composite log-likelihood for fitting generalized Pareto distributions to exceedances over a threshold. Some theory demonstrates the asymptotic advantages of using weights in the special case when the shape parameter of the limiting generalized Pareto distribution is known to be 0. These theoretical results show clear parallels to results for choosing weight functions and bandwidths for kernel density estimation. Methods for extending this approach to observations that are not independent and identically distributed are described and an analysis to daily precipitation data in New York City provided. Perhaps the most important practical finding is that including weights in the composite log-likelihood can reduce the sensitivity of estimates to small changes in the threshold.

Keywords Extreme value theory · Generalized Pareto distribution · Order statistics · Kernel density estimation

Mathematics Subject Classification (2010) 62G32 · 62G30

M. L. Stein
Department of Statistics
Rutgers University
Piscataway, NJ
E-mail: ms2870@stat.rutgers.edu
https://orcid.org/0000-0002-2059-2790

1 Introduction

Estimating extreme quantiles of a distribution provides a striking case of the bias-variance tradeoff that is ubiquitous throughout statistics. Specifically, from a sample of size n from some distribution, quantiles that are not too extreme can be estimated with little bias under weak assumptions on the smoothness of the underlying distribution (see, e.g., Falk (1985)). However, without additional assumptions, it is difficult to say something useful about quantiles like $1-\delta_n$, where δ_n is comparable to or, worse, much smaller than n^{-1}. Alternatively, assuming that the observations come from some low-dimensional parametric family can yield good estimates of even extreme quantiles if the family includes the true distribution but at the risk of severe bias if the truth is not in this family. Thus, it is common practice to appeal to extreme value theory, which shows that for a broad range of distributions, their upper tails can be well approximated by a generalized Pareto distribution (GPD), whose survival function is given by

$$G_\theta(x) = \left(1 + \frac{\xi(x-\mu)_+}{\sigma}\right)_+^{-1/\xi}, \tag{1}$$

where $\theta = (\mu, \sigma, \xi)$ with μ and ξ real, σ positive and y_+ is the positive part of y. For $\xi = 0$, this formula is interpreted as the limit as $\xi \to 0$, given by $\exp\{-(x-\mu)_+/\sigma\}$. This result can be used to estimate extreme quantiles by, for example, using the $n-j$'th order statistic to estimate the $1-(j+1)/(n+1)$ quantile of the distribution and then treat exceedances above this order statistic as iid (independent and identically distributed) from a GPD with $\mu = 0$. Typically, j is chosen to be large but small relative to n.

Using this formulation does not make the bias-variance problem go away. Indeed, the situation is somewhat analogous to nonparametric density estimation, in which one compromises between assuming a finite-parameter model for the density and assuming nothing about it by assuming the density is, say, twice differentiable, which allows one to obtain rates of convergence for estimates of the density at any fixed x for which the density is positive (Sheather, 2004). For example, one can just use a boxcar kernel estimate of the density of x, given by the fraction of the observations within some distance b of x divided by $2b$. This bandwidth parameter b plays a similar role as j, the number of tail observations used to estimate the parameters of the GPD. It is well-known that density estimates with modestly lower asymptotic variance can be obtained by using a kernel that decreases smoothly as observations get farther from x (Epanechnikov, 1969). Furthermore, if one is willing to use a kernel function that is negative at some distances, it is possible to obtain estimates with a faster rate of convergence than can be achieved with any nonnegative kernel (Jones and Signorini, 1997; Sheather, 2004). This work develops an analogous approach to smoothly downweighting the influence of less extreme observations for fitting GPDs to the tails of a sample from some distribution.

Section 2 describes the basic methodology, which exploits the elementary fact that the order statistics from an iid sample form a Markov chain. This

result allows one to write down a weighted composite log-likelihood (WCL) for the parameters of the GPD distribution that includes the usual log-likelihood function when the weights are all 1; see Varin et al. (2011) for a general review of weighted composite log-likelihood methods. When the shape parameter ξ of the GPD is assumed to be 0, so that the GPD reduces to an Exponential distribution, then the maximizing value for the scale parameter σ of the WCL has a closed form expression, which then also yields a closed form expression for the corresponding estimated quantiles. This explicit expression for the estimated quantiles is used in Section 3 to show that under a standard condition on the second-order tail behavior of the distribution and allowing j to grow at an appropriate rate as n increases, the estimated quantiles are asymptotically normal with asymptotic bias related to this second-order tail behavior. These results allow one to compare mean squared errors for different weighting functions and, assuming an exponent appearing in the expression for the second-order behavior of the underlying distribution is known, an asymptotically optimal nonnegative weighting function can be obtained. Furthermore, if this exponent is known, it is possible to eliminate the leading term in the bias of the quantile estimates by using a weighting function that is negative on part of its domain. Simulation results in Section 4 support these results and also show benefits to weighting in the more realistic setting where the shape parameter is not assumed known. In addition to yielding somewhat better quantile estimates if j is chosen well, using weights can substantially reduce the volatility of the estimated quantiles as j varies.

Extreme value methods are often applied to observations that are not independent or not identically distributed or neither. The Markov property of order statistics only applies when the observations are iid. In a brief simulation study, Section 5 examines how dependence can effect the expected values of gaps in order statistics and thus bias estimates based on the WCL. This bias can be substantially reduced by partitioning the observations into subsets made up of well-spaced observations. Furthermore, an additional simulation shows that, at least in some circumstances, biases in estimated quantiles due to dependence can be substantially smaller than biases due to exceedances from the underlying distribution not being exactly from a GPD. Section 6 considers independent but not identically distributed observations and writes down two conditional distributions based on order statistics that could be used to obtain WCLs in this setting. Because the order statistics no longer form a sufficient statistic when the observations are not iid, Section 6 proposes a simple approach based on quantile regressions to preprocess the observations so that they are more nearly iid before reducing them to order statistics. In Section 7, this approach of applying WCL to preprocessed observations is compared to estimates based on the ordinary likelihood for over 150 years of daily rainfall data in New York City. Specifically, for a range of small probabilities δ, a seasonally varying threshold function was estimated using quantile regression for the $1 - \delta$ quantile and then exceedances beyond this threshold function were fit using a GPD with seasonally varying scale parameter. For the WCL, the differences between quantile regressions at the $1 - \delta/2$ and $1 - \delta$ quantiles

were used to preprocess the exceedances to make them more nearly identically distributed. These procedures were fitted to roughly half of the data and then evaluated using the other half of the data. Measured by a log-likelihood for the testing data censored at cutoffs of 2 or 3 inches, the weighted approach has the advantage that its results vary much less erratically with the choice of δ. When δ is chosen well, both approaches yield good estimates of seasonally varying extreme quantiles for daily rainfall in the testing data.

Section 8 discusses further possible applications of WCLs based on order statistics, both to extremes and more generally, as well as other issues such as how to choose thresholds in practice. Section 9 contains the proof of Theorem 1 from Section 3.

2 Methodology

Let us first consider the general setting of fitting a parametric model to iid observations. Suppose $X_1, \ldots, X_n$ are iid from some density f_θ from a family of densities indexed by the parameter θ with corresponding cumulative distribution function F_θ and survival function $S_\theta = 1 - F_\theta$. Writing $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ for the corresponding order statistics, then because the order statistics form a Markov chain (David and Nagaraja, 2003)[Section 2.5], the log-likelihood can be written as (ignoring an additive constant)

$$\log f_\theta(X_{(1)}) + (n-1)\log S_\theta(X_{(1)}) + \sum_{k=1}^{n-1} \{(k-1)\log S_\theta(X_{(n-k+1)}) + \log f_\theta(X_{(n-k+1)}) - k \log S_\theta(X_{(n-k)})\}.$$

Of course, it would be pointless to use this form of the log-likelihood function if F_θ included the true distribution. However, if one were more interested in some parts of the distribution than others and were concerned that the parametric model could be misspecified in ranges that were not of particular interest, then this form of the log-likelihood naturally allows weighting:

$$w^*\{\log f_\theta(X_{(1)}) + (n-1)\log S_\theta(X_{(1)})\} + \sum_{k=1}^{n-1} w_k\{(k-1)\log S_\theta(X_{(n-k+1)}) + \log f_\theta(X_{(n-k+1)}) - k \log S_\theta(X_{(n-k)})\} \tag{2}$$

for constants $w^*, w_1, \ldots, w_{n-1}$. The corresponding weighted score equations obtained by setting derivatives with respect to the components of θ to 0 are, under standard regularity conditions, unbiased estimating equations when F_θ includes the true distribution (Li and Babu, 2019)[Chapter 9].

When fitting the parameters of a generalized Pareto distribution to the largest order statistics, it is common to use the conditional log-likelihood of $X_{(n-j+1)}, \ldots, X_{(n)}$ given $X_{(n-j)}$, given by

$$\sum_{k=1}^{j} \{(k-1)\log S_\theta(X_{(n-k+1)}) + \log f_\theta(X_{(n-k+1)}) - k \log S_\theta(X_{(n-k)})\}.$$

Similar to (2), we can gradually downweight the influence of less extreme observations:

$$\ell(\theta; w) = \sum_{k=1}^{j} w_k \{(k-1)\log S_\theta(X_{(n-k+1)}) + \log f_\theta(X_{(n-k+1)}) - k \log S_\theta(X_{(n-k)})\}. \tag{3}$$

The idea of a local likelihood for dependent observations goes back to at least Rao (1970) in the context of regularly observed time series. Anderes and Stein (2011) used a similar approach to fit the parameters of a stationary covariance function to spatial data in the neighborhood of some location y_0 when the actual process is only "locally stationary" in an appropriate sense. In contrast with (3), observations in these works were ordered by differences in time or distance from a point of interest rather than by the value of the process. Lawrance (1990) suggests adding weights to the log-likelihood function of the original observations when fitting extreme value distributions as a method of evaluating the influence of one or more observations, but this is a rather different goal than downweighting less extreme observations in parameter estimation.

If, as is common, we assume that exceedances of $X_{(n-j)}$ are from a GPD with $\mu = 0$, then defining $Y_i = X_{(n-j+i)} - X_{(n-j)}$, so $Y_0 = 0$,

$$\ell(\theta; w) = \sum_{k=1}^{j} w_k \left\{ -\left(\frac{k}{\xi} - 1\right) \log\left(1 + \frac{\xi Y_{j-k+1}}{\sigma}\right) - \log\sigma + \frac{k}{\xi} \log\left(1 + \frac{\xi Y_{j-k}}{\sigma}\right) \right\}. \tag{4}$$

When $\xi = 0$ is known, this approach yields closed-form expressions for the estimate σ. The WCL is given by

$$-\sum_{k=1}^{j} w_k \left\{ \frac{k(Y_{j-k+1} - Y_{j-k})}{\sigma} - \log\sigma \right\}. \tag{5}$$

and maximizing (5) with respect to σ yields

$$\begin{aligned} \hat{\sigma} &= W_j^{-1} \sum_{k=1}^{j} w_k k (Y_{j-k+1} - Y_{j-k}) \\ &= W_j^{-1} \sum_{k=1}^{j} \{k w_k - (k-1) w_{k-1}\} Y_{j-k+1}, \end{aligned} \tag{6}$$

where $W_j = \sum_{k=1}^{j} w_k$ and w_0 can be set to 0. If the X_i's are actually from an Exponential distribution, then $\hat{\sigma}$ in (6) is unbiased for σ, which follows from Rényi's representation for order statistics of an Exponential distribution (Rényi, 1953).

For $p \geqslant (n-j)/(n+1)$, the corresponding estimated p'th quantile is

$$\begin{aligned}\hat{Q}(p) &= X_{(n-j)} - \hat{\sigma}\log\left((1-p)\frac{n+1}{j+1}\right) \\ &= X_{(n-j)}\left\{1 + \frac{jw_j}{W_j}\log\left((1-p)\frac{n+1}{j+1}\right)\right\} \\ &\quad - W_j^{-1}\log\left((1-p)\frac{n+1}{j+1}\right)\sum_{k=1}^{j}\{kw_k - (k-1)w_{k-1}\}X_{(n-k+1)}. \quad (7)\end{aligned}$$

This last form for the estimated quantiles facilitates the study of their asymptotic properties when $\xi = 0$ is treated as known in Section 3.

The kernel-like form of (4) is reminiscent of kernel estimates for ξ proposed in the literature by Csorgo et al. (1985), based on adding weights to the classical Hill estimator of ξ when $\xi > 0$. Further study of this weighted Hill estimator is provided in Caeiro et al. (2019) and the references therein. Groeneboom et al. (2003) describe a kernel estimator for ξ that works for all real ξ, although their functional forms are not nearly as simple as for the weighted Hill estimator. The estimates in all of these works are in terms of the logarithms of the largest order statistics and thus only make sense for positive random variables. The method described here does not require positive observations and, indeed, is location-invariant for all ξ. As noted by Beirlant et al. (2012), many (but certainly not all) methods for estimating extremes lack location-invariance. Whether location invariance is always desirable is not a simple question. In particular, for intrinsically positive quantities, 0 is a special value, so location invariance may sometimes be inappropriate, although scale invariance would still generally be desirable. Perhaps more importantly, it is unclear how the kernel methods in Groeneboom et al. (2003) could be extended to observations that are not identically distributed, whereas the WCL based on order statistics can be applied in this setting (Section 6), if perhaps not totally satisfactorily.

As in the case of density estimation, one would generally want the w_k's to decrease smoothly and tend to 0 as k approaches j. This can be accomplished by setting, for $k > 0$, $w_k = \omega((k-1)/j)$ for some positive and continuous decreasing function ω on $[0,1]$ with $\omega(1) = 0$. All of the weight functions ω used in the simulations and the data analysis satisfy $\omega(1) = 0$.

3 Asymptotic theory

When fitting GPDs to the largest order statistics from a sample, we are implicitly assuming that the upper tail of the true distribution is in the domain of attraction of some element of the family. The class of distributions that satisfies this property is so broad that it is necessary to restrict what true distributions one wishes to consider in order to obtain results that would shed any light on how to pick the weights in the WCL (3). I only consider a narrow special case here, where the true distribution behaves like an Exponential

distribution in its upper tail and $\xi = 0$ is treated as known. Following the program in Drees (1998) or de Haan and Ferreira (2006), it should be possible to obtain results treating ξ as unknown, but even the limited setting studied here provides some insights into the choice of the weights and evidence for the asymptotic superiority of unequal weights.

Assume that $w_k = \omega((k-1)/j)$ for some sufficiently smooth function $\omega(\cdot)$ and that $\int_0^1 \omega(t)\,dt = 1$ to fix the normalization of ω. To motivate the conditions on the true distribution used in Theorem 1 below, for some $\lambda \in (0,1)$ and some positive α and γ, consider a mixture of two Exponential distributions, whose survival function is

$$S(x) = \lambda e^{-\alpha x} + (1-\lambda)e^{-\alpha(\gamma+1)x} \tag{8}$$

for $x \geqslant 0$. Since only the asymptotic behavior of S in its upper tail will matter for the results here, we will consider the somewhat more general setting in which

$$S(x) = a_1 e^{-\alpha x} + a_2 e^{-\alpha(\gamma+1)x} + o\big(e^{-\alpha(\gamma+1)x}\big) \tag{9}$$

as $x \to \infty$. By straightforward computations, the corresponding quantile function, Q, satisfies

$$Q(1-\epsilon) = -\frac{1}{\alpha}\log\epsilon + \frac{1}{\alpha}\log a_1 + \frac{a_2}{\alpha a_1^{\gamma+1}}\epsilon^\gamma + o(\epsilon^\gamma) \tag{10}$$

as $\epsilon \downarrow 0$. In fact, Theorem 1 below only requires that Q satisfy (10). For example, for a logistic distribution, whose survival function S is of the form $1/(1+e^{(x-\mu)/\sigma})$, we have $\alpha = \sigma^{-1}$, $\gamma = 1$, $a_1 = e^{\mu/\sigma}$ and $a_2 = -e^{2\mu/\sigma}$ in (10).

Under (10),

$$\frac{Q(1-tx) - Q(1-t)}{1/\alpha} = -\log x + \frac{a_2 t^\gamma}{a_1^{\gamma+1}}(x^\gamma - 1) + R(t,x), \tag{11}$$

where, for any fixed $x > 0$, we have $R(t,x) = o(t^\gamma)$ as $t \downarrow 0$. Defining $\Psi(x) = x^\gamma - 1$ and $\Phi(t) = a_2 t^\gamma / a_1^{\gamma+1}$, it follows that Condition 1 in Drees (1998) is satisfied. Thus, we can use Theorem 2.1 from Drees (1998) to find the limiting asymptotic distribution of the estimated quantiles in the important case where j grows with n in a way that the bias and standard deviation of these quantiles are of the same order of magnitude.

Theorem 1 *Suppose $X_1, X_2, \ldots$ are iid with distribution satisfying (10) and, for finite C, the sequence of positive integers j_n satisfies*

$$\lim_{n\to\infty} \sqrt{j_n}\Phi(j_n/n) = C. \tag{12}$$

Assume ω has a bounded second derivative on $[0,1]$ and $\int_0^1 \omega(t)\,dt = 1$. Then for a positive sequence $\delta_n = o(j_n/n)$ and $\hat{Q}$ defined as in (7),

$$\frac{\alpha\sqrt{j_n}}{\log(j_n/(n\delta_n))}\big(\hat{Q}(1-\delta_n) - Q(1-\delta_n)\big) \to \mathrm{N}\left(-C\int_0^1 t^\gamma \omega(t)\,dt, \int_0^1 \omega(t)^2\,dt\right) \tag{13}$$

in distribution.

The proof of this result is given in Section 9. For working out the implications of this result, it is helpful to replace j_n by its asymptotic approximation (41), so that Theorem 1 implies

$$\alpha\left(\frac{a_1^{\gamma+1}}{a_2}\right)^{1/(2\gamma+1)} \frac{n^{\gamma/(2\gamma+1)}}{\frac{2\gamma}{2\gamma+1}\log n - \log(n\delta_n)}\big(\hat{Q}(1-\delta_n) - Q(1-\delta_n)\big)$$
$$\to \mathrm{N}\left(-C^{2\gamma/(2\gamma+1)}\int_0^1 t^\gamma\omega(t)\,dt, C^{-2/(2\gamma+1)}\int_0^1 \omega(t)^2\,dt\right) \tag{14}$$

in distribution. Writing C_γ for the value of C that minimizes the second moment of this limiting distribution, we have

$$C_\gamma = \frac{\big\{\int_0^1 \omega(t)^2\,dt\big\}^{1/2}}{(2\gamma)^{1/2}\left|\int_0^1 t^\gamma\omega(t)\,dt\right|}, \tag{15}$$

assuming the denominator is not 0. Defining

$$I_\gamma(\omega) = \left(\int_0^1 \omega(t)^2\,dt\right)^{2\gamma/(2\gamma+1)}\left|\int_0^1 t^\gamma\omega(t)\,dt\right|^{2/(2\gamma+1)}, \tag{16}$$

the corresponding expression for the minimized second moment of the asymptotic distribution of $\hat{Q}(1-\delta_n) - Q(1-\delta_n)$ is

$$\frac{2\gamma+1}{\alpha^2}\left(\frac{a_2}{a_1^{\gamma+1}(2\gamma)^\gamma}\right)^{2/(2\gamma+1)} \cdot \frac{I_\gamma(\omega)\left(\frac{2\gamma}{2\gamma+1}\log n - \log(n\delta_n)\right)^2}{n^{2\gamma/(2\gamma+1)}}, \tag{17}$$

which is attained when j_n satisfies

$$j_n \sim \left(\frac{a_1^{2\gamma+2}\int_0^1 \omega(t)^2\,dt}{2a_2^2\gamma\big\{\int_0^1 t^\gamma\omega(t)\,dt\big\}^2}\right)^{1/(2\gamma+1)} n^{2\gamma/(2\gamma+1)}. \tag{18}$$

Of course, (18) is not useful for selecting j_n since it requires knowing $a_1^{\gamma+1}/a_2$ in addition to γ. However, if there were some basis for believing that γ takes on some particular value such as 1 or 2, then (17) could be useful for selecting ω because $I_\gamma(\omega)$ depends only on γ.

If we allow ω to take on negative values, we can make $\int_0^1 t^\gamma\omega(t)\,dt = 0$ for any given positive value of γ. Section 4 shows some limited results for such a kernel; there is a substantial reduction in bias from using a kernel that makes $\int_0^1 t^\gamma\omega(t)\,dt = 0$ for the true value of γ, but the improvement in mean squared error is less dramatic. Restricting ω to be nonnegative, we can, for any given γ, seek the nonnegative function ω that integrates to 1 on $(0,1)$ and minimizes $I_\gamma(\omega)$. This problem is just a one-sided version of Lemma 18 in Devroye and Györfi (1985)[Chapter 5], so an optimal nonnegative ω is given by

$$\omega_\gamma(t) = \frac{\gamma+1}{\gamma}(1-t^\gamma). \tag{19}$$

This minimizer is not unique since, for any $c > 1$, the function $c\omega(ct)$ also minimizes the functional I_γ.

If $n\delta_n = O(1)$, then the term $\log(n\delta_n)$ can be dropped from (14) and the result still holds, so that the actual quantile being estimated affects neither the large-sample bias nor variance to this degree of approximation. Examining the proof of Theorem 1, we can see that (13) drops out terms that are smaller than the included terms by a factor of only $\log j_n$ and, thus, may not be very accurate unless j_n is quite large. Assuming an appropriately stronger condition on j_n than (12), Equations (39)–(42) suggest that a better approximation than implied by Theorem 1 might be obtained by adding $C/(\alpha\sqrt{j_n})$ to the mean of $\hat{Q}(1-\delta_n) - Q(1-\delta_n)$, or that

$$\begin{aligned}&\hat{Q}(1-\delta_n) - Q(1-\delta_n)\\ &\approx \mathrm{N}\left(-\frac{C}{\alpha\sqrt{j_n}}\log\frac{n\delta_n}{j_n}\int_0^1 t^\gamma\omega(t)\,dt + \frac{C}{\alpha\sqrt{j_n}}, \frac{C^2}{\alpha^2 j_n}\log^2\frac{n\delta_n}{j_n}\int_0^1\omega(t)^2\,dt\right).\end{aligned} \tag{20}$$

Proving that this result yields a better approximation than Theorem 1 would require sharper control of the error in the Brownian motion approximation used in Section 9.

In the case of a mixture of two Exponential distributions, it is possible to show directly that including the $C/(\alpha\sqrt{j_n})$ term provides a sharper approximation to the bias of $\hat{Q}(1-\delta_n)$. Because this result is so narrow, I just outline the argument. Suppose that $\sqrt{j_n}\Phi(j_n/n) = C + O(j_n^{-\epsilon})$ for some $\epsilon > 0$. Note that, from (6), we can write

$$\hat{\sigma} = W_{j_n}^{-1}\sum_{k=1}^{j_n} w_k k(X_{(n-k+1)} - X_{(n-k)}). \tag{21}$$

Since $Z_i = -\log(1-F(X_i))$ follows a standard Exponential distribution, for $k = 1, ..., n$, the gaps in the order statistics $Z_{(n-k+1)} - Z_{(n-k)}$ (where $Z_{(0)} = 0$) are independent Exponentials with $E(Z_{(n-k+1)} - Z_{(n-k)}) = k^{-1}$ (Rényi, 1953). Defining $H(z) = Q(1-e^{-z})$ so that $Z_i = H^{-1}(X_i)$ and taking a second order Taylor series, we have

$$\begin{aligned}&X_{(n-k+1)} - X_{(n-k)}\\ &= H'(EZ_{(n-k)})(Z_{(n-k+1)} - Z_{(n-k)})\\ &\quad + \frac{1}{2}H''(EZ_{(n-k)})\left\{(Z_{(n-k+1)} - EZ_{(n-k)})^2 - (Z_{(n-k)} - EZ_{(n-k)})^2\right\}\end{aligned}$$

plus a remainder that can be shown to contribute $o\left(j_n^{-1/2}\right)$ to the expectation of $\hat{Q}(1-\delta_n)$. Taking expectations yields the approximation

$$E(X_{(n-k+1)} - X_{(n-k)}) \approx H'(EZ_{(n-k)})\frac{1}{k} + H''(EZ_{(n-k)})\frac{1}{k^2}. \tag{22}$$

From Olver et al. (2010)[(2.10.8] we get

$$EZ_{(n-k)} = \sum_{\ell=k+1}^{n} \frac{1}{\ell} = \log \frac{n}{k} + \frac{1}{2n} - \frac{1}{2k} + O\left(\frac{1}{k^2}\right). \tag{23}$$

Applying (21)–(23) to (7), dropping lower order terms and approximating sums by integrals yields the desired result on $E\hat{Q}(1-\delta_n)$. Section 4 gives a numerical example showing that this approximation is an improvement over the mean from Theorem 1.

Davison and Smith (1990)[Section 10] note a similarity between threshold selection, or nearly equivalently, the choice of j, and bandwidth selection in kernel density estimation. By allowing weights in the composite log-likelihood, the similarity between fitting GPDs to exceedances and kernel density estimation is made even stronger. In particular, when considering mean squared error of kernel density estimates for densities with two derivatives, asymptotic results that look quite a lot like (14)–(18) with $\gamma = 2$ occur (Rosenblatt, 1971), although without the $\log n$ terms in (14) and (17). Values of γ less than 2 can arise in the asymptotics of kernel density estimation for densities not possessing a second derivative. For example, Devroye and Györfi (1985)[Chapter 5, Section 7] essentially shows that $\gamma = 1$ is the appropriate value to use when evaluating kernel density estimates of a uniform density, which does not even possess one derivative at its endpoints. In the setting studied here, γ represents how well the Exponential distribution approximates the upper tail of the true distribution, with larger γ corresponding to better approximation. The result that allowing ω to take on negative values can eliminate the leading term in a bias expansion is well-known in the kernel density estimation literature (Jones and Signorini, 1997).

In the case where $\xi = 0$ is known, Theorem 1 can be used to obtain a relatively consistent estimate of the variance of $\hat{Q}(1-\delta_n)$ by replacing α in (13) by the maximized WCL estimate. However, unless at least some lower bound on the second-order parameter γ is assumed and j_n is deliberately chosen to increase sufficiently slowly to guarantee $C = 0$ in (12), then it is not clear how one might obtain an asymptotically valid confidence interval for $\hat{Q}(1-\delta_n)$. Again, there is a clear similarity to kernel density estimation, for which the limiting bias of the density estimate is non-negligible if the bandwidth is chosen to minimize mean squared error (Sheather, 2004). In many problems in extremes, the observations form a time series, in which case the variance approximation is also likely to be suspect because of the effect of dependence of neighboring observations. Section 5 considers extending the WCL to stationary processes and shows that the bias of point estimates due to ignoring temporal dependence in the WCL can be minimal in at least some circumstances in which extremal dependence is very strong. However, the same cannot be expected to hold for the variance of the estimates, in which case, some kind of resampling approach (Gomes and Neves, 2015) may be a better option than asymptotic approximation.

4 Simulation study

This section describes some results from a limited simulation study. To investigate how well the asymptotic results in the previous section work, I first consider parameter and quantile estimation when $\xi = 0$ is known and the true distribution satisfies (9). The simulations show that the heuristic approximation in (20) gives a better approximation to the bias of quantile estimates than (13) in Theorem 1. I then consider some simulations from heavy-tailed distributions and ξ is treated as unknown. In both settings, the WCL with weight function ω_1 as defined in (19) performs slightly but distinctly better than the ordinary likelihood. Perhaps more importantly in practice, the estimate of ξ varies much less erratically with the bandwidth parameter j when using a weight function ω for which $\omega(1) = 0$.

Figure 1 shows results from a simulation study in which $\xi = 0$ is treated as known. The simulation is made up of 5000 replications of iid samples of size 6400 from a mixture of Exponentials with $\alpha = \gamma = 1$ and $\lambda = 0.5$ in (8). Large sample sizes and large numbers of simulations aid in seeing some of the fairly small differences between the results for different weight functions. The three weight functions shown are: constant weights, ω_1 in (19), which is asymptotically optimal in this example among nonnegative weight functions, and $\tilde{\omega}(t) = 6 - 18t + 12t^2$, for which $\int_0^1 t\tilde{\omega}(t)\,dt = 0$, so that this kernel gives a lower order asymptotic bias than the other kernels when $\gamma = 1$. This particular functional form for $\tilde{\omega}$ was chosen because it is the unique quandratic polynomial P that satisfies $\int_0^1 P(t)\,dt = 1$, $P(1) = 0$ and $\int_0^1 tP(t)\,dt = 0$. The left panel shows the mean of the estimates of the scale parameter σ as a function of the bandwidth j, which should be compared to the true scale parameter of the upper tail, given by $\alpha^{-1} = 1$. We see that, as expected, for any given value of j, the weight function ω_1 gives noticeably smaller bias than constant weights and $\tilde{\omega}$ has much lower bias than either of them. Of course, for a given value of j, the constant weight function yields an estimate with the lowest variance. When minimizing the mean squared error over j for constant weights, the minimizing j is 319 with mean squared error of 0.00486, for ω_1, the minimizing j is 469 with mean squared error of 0.00453, and for $\tilde{\omega}$, the minimizing j is much larger, 1972, with mean squared error 0.00350. Another possible advantage of varying weights can be seen in the right panel of Figure 1, which shows that the mean squared error changes more slowly with the bandwidth j for ω_1 and especially $\tilde{\omega}$ than for constant weights. Thus, we might hope that selecting j will be less critical when using an appropriate WCL, although even empirical verification of such a claim would depend on having a unified approach for selecting j across some class of weight functions.

Figure 2 compares empirical and asymptotic biases and variances for estimated quantiles from these same simulations. Write δ_n in the form δ/n to make it easier to see to what extent the quantile being estimated is an extrapolation outside the range of the observations. As we should expect, the bias in $\hat{Q}(1 - \delta/n)$ increases as δ decreases. Furthermore, as the asymptotics indicate, the change in bias is essentially linear in $\log \delta$. Theorem 1 gets the

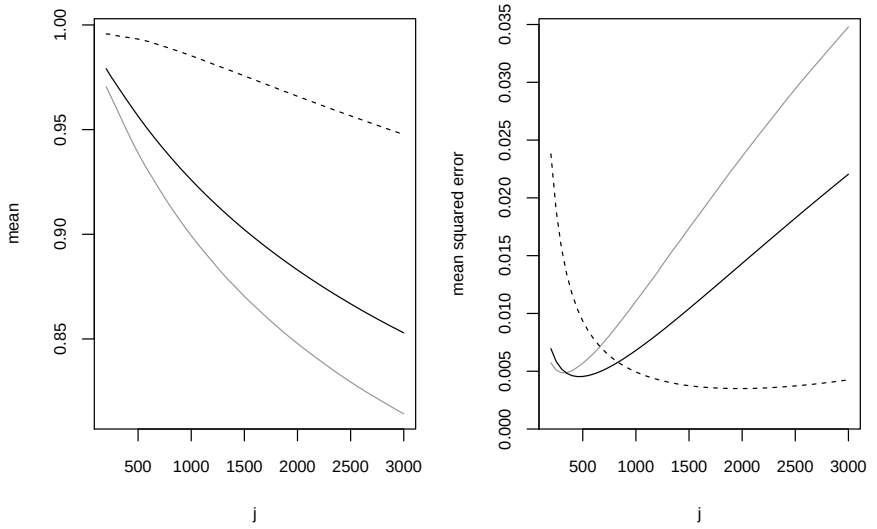

Fig. 1 Simulated means and mean squared errors for estimates of σ (truth equals 1) when $\xi = 0$ is known based on 6400 observations from a mixture of Exponentials. The horizontal axis gives j, the number of observations used to estimate σ. Solid gray line corresponds to constant weights, solid black to $\gamma = 1$ in (19) and dashed black to quadratic weight function that removes the leading term in the bias.

slope of this changing bias very accurately, but with a substantial error in the intercept. Using the correction suggested by (20) reduces the error in this intercept. Theorem 1 accurately captures the variance of $\hat{Q}(1 - \delta/n)$ for the wide range of δ values shown.

Of course, the more important practical setting is when ξ is treated as unknown. Figure 3 shows mean squared errors for estimates of ξ based on samples of size $n = 6400$ for two distributions in which the true value of ξ is $1/3$, so fairly heavy-tailed distributions. The left plot is for a mixture of two standard ($\mu = 0$ and $\sigma = 1$) GPDs, where the first component has mixing probability 0.1 and $\xi = 1/3$ and the second component has mixing probability 0.9 and $\xi = 1/6$. The right plot is for the cumulative distribution function $T_3(x)^2$, where T_ν is the cumulative distribution function of a t distribution on ν degrees of freedom. Both of these distributions satisfy Condition 1 in Drees (1998): there exist constants c_1 and c_2 such that

$$\frac{Q(1 - tx) - Q(1 - t)}{c_1 t^{-1/\xi}} = \frac{x^{-\xi} - 1}{\xi} + c_2 t^{1+\gamma}(x^\gamma - 1) + R(t, x), \qquad (24)$$

where $\xi = 1/3$, $\gamma = 1$ and $R(t, x) = o(t^{1+\gamma})$ as $\downarrow 0$ for all $x > 0$. Note that the power of x in the second order term is γ, which equals 1 in both simulated distributions. Thus, as for the mixture of Exponentials simulations, we might expect ω_1 as defined in (19) would perform well among nonnegative weight functions for large sample sizes and, indeed, Figure 3 shows ω_1 slightly outperforms ω_2, which in turn slightly outperforms the constant weight function.

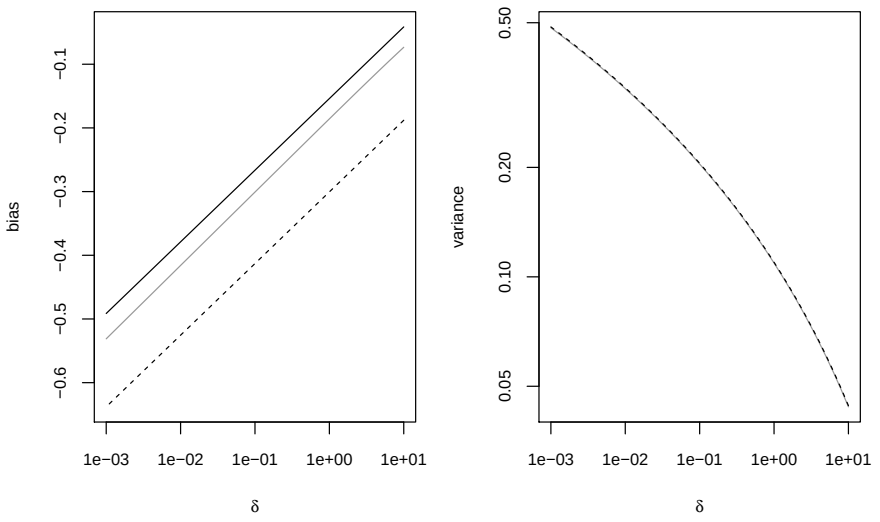

Fig. 2 Bias and variance of $\hat{Q}(1-\delta/n)$ for sample size $n = 6400$ when $\xi = 0$ is known. Left panel shows empirical bias (solid gray curvex), asymptotic bias using Theorem 1 (dashed black curve) and approximate bias using (20) (solid black curve). Right panel shows empirical variance (solid gray curve) and asymptotic variance using Theorem 1 (dashed black curve).

Despite using 5000 simulations, the mean squared errors for the constant weight function in Figure 3 have noticeable fluctuations in j, especially for the mixture distribution, suggesting that there is considerable variation in j for estimates of ξ when using a constant weight function. These fluctuations for the mixture model are explicitly shown in Figure 4 for the first 2 of the 5000 simulations run. For both simulations and all three weight functions, while the estimates of ξ change substantially with j, the changes are fairly smooth for ω_1 and ω_2, but are quite erratic for constant weights. It is not surprising that for a weight function $\omega(t)$ that tends to 0 as $t \uparrow 1$, a small increase in j yields a small change in the estimate of ξ, since the new observations incorporated into the estimate would all get small weights. Plotting estimates of ξ as a function of threshold and selecting a threshold beyond which these estimates no longer show clear systematic variation gives one informal approach for selecting the threshold when fitting a GPD (Coles, 2001)[p. 83]. Applying this approach to these two simulations would seem to suggest that, for the weighted procedures, the threshold should be taken even larger than the largest value shown in these plots, corresponding to $j = 100$. However, as the left panel in Figure 3 shows, the mean squared error for these weighted estimates are minimized when $j > 600$, so this graphical method for threshold selection may not be suitable for WCL estimates. In particular, the reduced fluctuations with small changes in j may make it too easy to see trends in $\hat{\xi}$, which could lead to choosing j too small if one expects to see no clear systematic trend in $\hat{\xi}$ beyond the selected value of j.

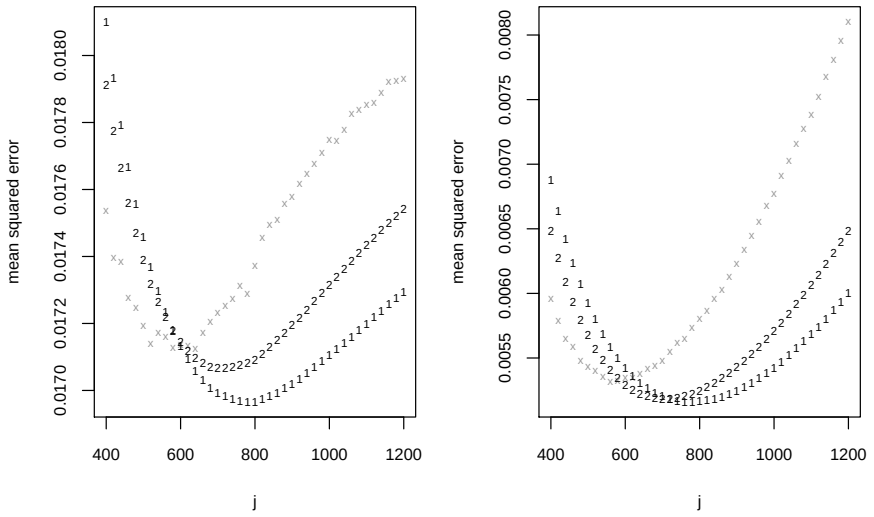

Fig. 3 Simulated mean squared errors for estimates of ξ based on 6400 observations as a function of j, the number of observations used to estimate ξ for (left plot) a mixture of two GPDs and (right plot) a distribution with cdf given by $T_3(x)^2$, where T_ν is the cdf of a t distribution on ν degrees of freedom. Plotting symbols indicate weight function, where gray ×'s indicate constant weights and numbers correspond to γ in (19).

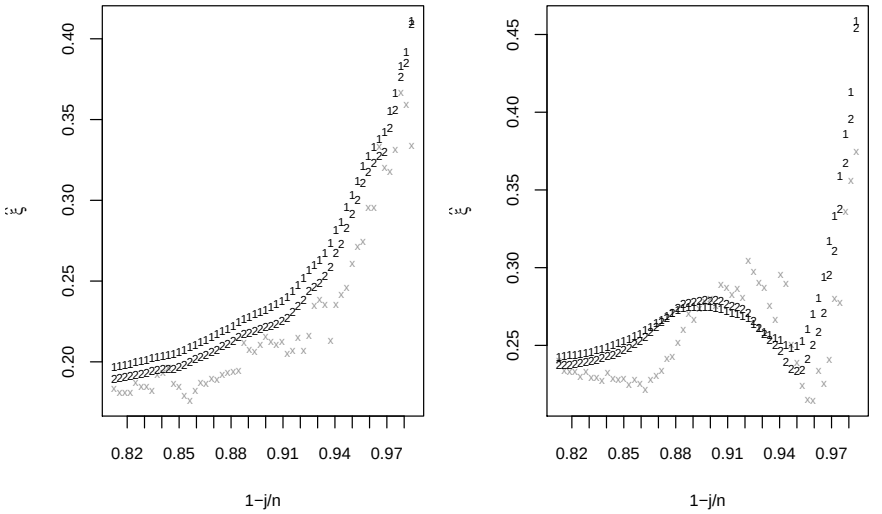

Fig. 4 Estimates of ξ from two simulations of 6400 observations from a mixture of two GPDs (true value of ξ is 1/3) as a function of the fraction of observations below the threshold. Plotting symbols indicate weight function, where gray ×'s indicate constant weights and numbers correspond to γ in (19).

5 Observations from a stationary process

When observations are identically distributed but not independent and we fit a parametric model to their marginal distribution acting as if they are independent, then the gradient of the resulting log-likelihood function still yields unbiased estimating equations. This property does not generally carry over to the WCLs in (2) and (3). However, we might hope that if the sample size is sufficiently large and the observations are not too dependent, the resulting biases might be small. For a stationary time series, declustering methods (Ferro and Segers, 2003) can reduce the impact of this dependence, but they can have their own problems with bias when the goal is estimation of marginal quantiles (Fawcett and Walshaw, 2007) and there are other approaches for addressing the impact of serial dependence on uncertainty quantification for extreme quantiles (Fawcett and Walshaw, 2012). Of course, if the dependence of extremes is of specific interest, then an appropriate data analysis will have to extend beyond estimating marginal quantiles.

If the dependence is negligible after a few lags, it should be possible to reduce any effects of this dependence on the bias in using a maximizer of (4) to estimate θ by splitting up the time series into k series with gaps of size k between each observation. For example, to estimate extreme upper quantiles using a GPD, first select a value j and set $\hat{Q}(1-(j+1)/(n+1)) = \hat{\mu} = X_{(n-j)}$ as before. Next, split the time series into m subseries, with the ℓ'th subseries $\mathbf{X}_\ell = (X_\ell, X_{\ell+m}, \ldots, X_{\ell+r_\ell m})$, where r_ℓ is chosen so that $\ell + r_\ell m \leqslant n < \ell + (r_\ell + 1)m$. Define $X_{(1,\ell)} \leqslant \cdots \leqslant X_{(r_\ell,\ell)}$ to be the ordered values in $\mathbf{X}_\ell$ and select $j(\ell)$ to satisfy $X_{(r_\ell - j(\ell),\ell)} \leqslant \hat{\mu} < X_{(r_\ell - j(\ell)+1,\ell)}$, where $X_{(0,\ell)} = -\infty$ and $X_{(r_\ell+1,\ell)} = +\infty$. Writing $Y_{k,\ell} = X_{r_\ell - j(\ell)+k,\ell} - \hat{\mu}$ for the exceedances of $\hat{\mu}$ and setting $Y_{0,\ell} = 0$, the resulting WCL for (σ, ξ) assuming observations in each subseries are independent is

$$\sum_{\ell=1}^{m} \sum_{k=1}^{j(\ell)} \omega\left(\frac{k-1}{j(\ell)}\right) \left\{ -\left(\frac{k}{\xi} - 1\right) \log\left(1 + \frac{\xi Y_{j(\ell)-k+1,\ell}}{\sigma}\right) - \log\sigma + \frac{k}{\xi} \log\left(1 + \frac{\xi Y_{j(\ell)-k,\ell}}{\sigma}\right) \right\}, \tag{25}$$

where a sum whose upper limit is less than its lower limit is understood to equal 0.

It appears difficult to study theoretically the bias in the estimating equations obtained by setting the gradient of (25) to 0 for dependent data, so here I show results from a small simulation study. I consider two stationary time series models in which the marginal distributions are Exponentials with scale parameter σ, so any bias that arises from using (25) is due to the dependence. To further simplify the problem, assume $\xi = 0$ and $\mu = 0$ are known and

$n/m = r$ is an integer, in which case, maximizing (25) yields

$$\hat{\sigma} = \frac{\sum_{\ell=1}^{m} \sum_{k=1}^{r} \omega\left(\frac{k-1}{r}\right) k(Y_{r-k+1,\ell,\ell} - Y_{r-k,\ell})}{m \sum_{k=1}^{r} \omega\left(\frac{k-1}{r}\right)}, \tag{26}$$

where $Y_{0,\ell} = 0$. Since $E(Y_{r-k+1,\ell,\ell} - Y_{r-k,\ell}) = \sigma/k$ when the observations are iid Exponential with scale parameter σ, we immediately see that this estimate is then unbiased. Using simulations for dependent Exponentials, we can then isolate the bias of these estimates that arises from using order statistics of dependent observations by considering $kE(Y_{j(\ell)-k+1,\ell,\ell} - Y_{(j(\ell)-k,\ell)})$ as k varies.

The first model considered is a chi-squared process with two degrees of freedom (Davies, 1987): the sum of squares of two independent stationary Gaussian AR(1) processes with mean 0, variance 0.5 and autocorrelation coefficient $\sqrt{\rho}$. The marginal distribution of the resulting process is standard Exponential and the lag m autocorrelation equal to ρ^m. The second model is the TEAR(1) process proposed in Lawrance and Lewis (1981) with correlation parameter $\rho \geqslant 0$, for which X_1 is a standard Exponential random variable and, for $n > 1$, X_n is defined recursively by $(1-\rho)E_n + B_{n-1}X_{n-1}$, where X_1, the B_n's and the E_n's are all independent, the E_n's follow a standard Exponential distribution and the B_n's follow a Bernoulli distribution with $P(B_n = 1) = \rho$. This process is stationary, the marginal distribution of each X_n is standard Exponential and the lag m autocorrelation is ρ^m. In the simulations, I set $\rho = 0.9$ for both processes to consider cases of strong dependence. As the upper left panel of Figure 5 shows, the realizations of the two processes look quite different. The first model is reversible in time whereas the second process clearly is not, with long periods of increase followed by sudden collapses, which occur when $E_n = 0$. From the point of view of extremes, the two processes are very different. The index of extremal dependence of consecutive observations (Chavez-Demoulin and Davison, 2012; Davison et al., 2013), defined as $\lim_{x\to\infty} P(X_{n+1} > x \mid X_n > x)$ (or as x tends to the upper limit of the distribution in the bounded case), is 1 for the first process, corresponding to asymptotically independent extremes, and equals $1-\rho$ for the second process.

Each process was simulated 100,000 times for sample sizes $n = 1600$ and 6400 and the average values for gaps of the order statistics were computed. Writing G_k for $X_{(n-k+1)} - X_{(n-k)}$, we know that $kEG_k = 1$ when $\rho = 0$. We also have $n^{-1}\sum_{k=1}^{n} kEG_k = 1$ for both models and all values of ρ, since the WCL then equals the exact log-likelihood under independence, which yields unbiased estimating equations even when the observations are dependent as long as the model for the marginal distribution of the observations is, as in the current case, correct. The upper right panel of Figure 5 shows the averages of kG_k over the 100,000 simulations for the two processes. These expected values are substantially less than 1 for $k = 1$ and then appear to decay exponentially towards 1 as k increases. Since $n^{-1}\sum_{k=1}^{n} kEG_k = 1$, some of the averages of kG_k must be greater than 1, but for $n = 6400$, all of these averages are well

below 1.02 for both processes, so the larger biases are all concentrated at small k. A perhaps unexpected result is that these averages show little dependence on n, to the extent that it is difficult to see the difference between the results for the two sample sizes for the TEAR(1) process. Since the value of j will tend to increase with n, this result suggests that bias induced by dependence will lessen as the sample size increases. Not surprisingly, the deviations from 1 are substantially larger for the TEAR(1) process, which has strong extremal dependence. The bottom two plots in Figure 5 show the average values of $k(Y_{r-k+1,\ell,\ell} - Y_{r-k,\ell})$ for $\ell = 1, 2$ and 4; for k small, we very nearly have $kE(Y_{r-k+1,\ell,\ell} - Y_{r-k,\ell}) = k\ell E(X_{(n-k\ell+1)} - X_{(n-k\ell)})$ for both processes. Further results (not shown) demonstrate that this approximation is accurate for larger values of r as well. Since one would generally use smaller values of j for smaller n, we might expect that biases caused by dependence would be more of an issue with smaller sample sizes.

Since the simulations of the previous section indicate that the exact form of the weights does not matter much, if one is concerned about the bias induced by dependence it may make sense to choose r fairly large as long as j/r is not too small. However, in practice, any bias due to dependence may be inconsequential compared to bias caused by the fact that the distribution of exceedances do not exactly follow a GPD. To demonstrate this phenomenon, consider a chi-squared process on 4 degrees of freedom: the sum of squares of four iid stationary Gaussian AR(1) processes with mean 0, variance 0.5 and autocorrelation coefficient $\sqrt{\rho}$. The marginal distribution of the resulting process is Gamma(2,1) and the lag m autocorrelation equals ρ^m. Since the correct asymptotic shape parameter is $\xi = 0$ in this case, I fit Exponential distributions to the upper order statistics for a range of j and r values using the linear weight function ω_1 as defined in (19). Figure 6 shows biases and standard deviations for $\rho = 0.9$, series of length 1500 and estimates of the $1 - 1/1500$ quantile, whose true value is 9.682. From Figure 5, we should expect the dependence to cause downward bias in the estimation of σ and, hence, downward bias in the estimate of extreme quantiles. Indeed, for the smaller values of j, the quantile estimates are downward biased for $r = 1$ and larger values of r correct this bias. However, even for $r = 1$, the biases eventually become positive for larger j so that increasing r makes this positive bias slightly larger. For any given value of j, the standard deviations of the estimates always increase with r, modestly for smaller j and minutely for larger j. Somewhat surprisingly, for given r, the standard deviations do not monotonically decrease with increasing j, being larger for $j = 40$ than for $j = 10$ or 20, although there is a clear decrease in the standard deviations with increasing j for $j \geqslant 80$. For the largest value of j shown of 640, in terms of the mean squared error, the standard deviations dominate the bias, so that among the combinations of j and r shown in Figure 6, $j = 640$ and $r = 1$ actually gives the smallest mean squared error, although larger values of r with $j = 640$ do only negligibly worse. Thus, we see that, at least in this case, the choice of j influences both the biases and standard deviations of the quantile estimates much more than the choice of r despite the strong autocorrelation in the process.

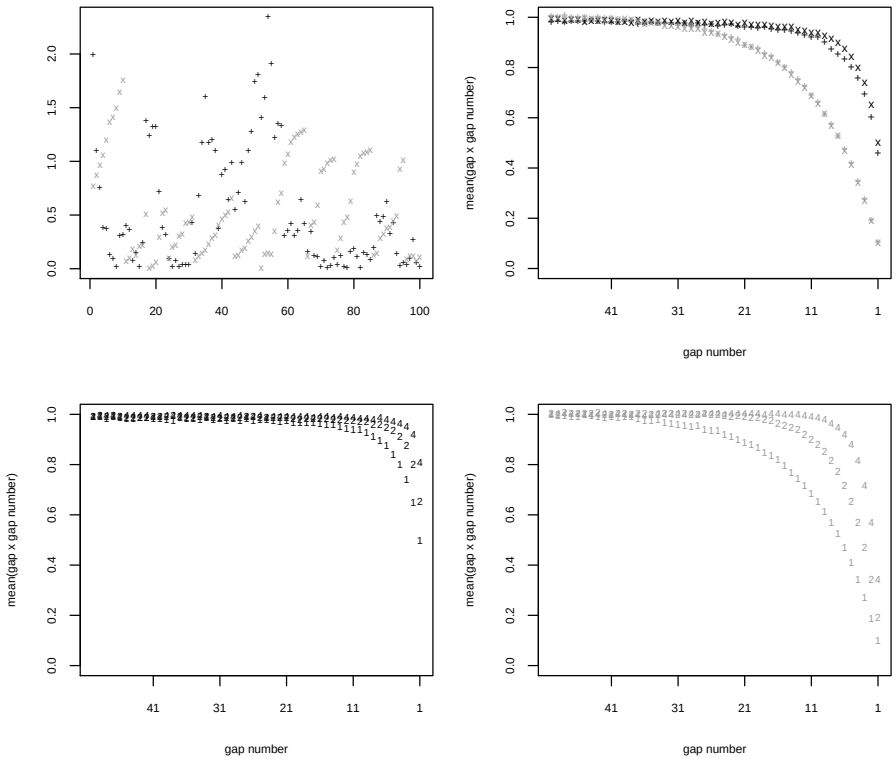

Fig. 5 Simulation results for stationary processes with mean 1 Exponential marginals and lag 1 autocorrelation 0.9. In all plots, black symbols correspond to chi-squared process and gray symbols to TEAR(1) process. Upper left panel: simulation of 100 time steps of each process. For a sample size n, upper right panel gives averages of $kX_{(n+1-k)}$ for $k = 50, \ldots, 1$, whose means would all be exactly 1 if the observations were independent. + corresponds to $n = 1600$ and × to $n = 6400$. For the TEAR(1) process, results are nearly identical for the two sample sizes, making it difficult to see both symbols. Lower two plots (left, chi-squared process, right, TEAR(1) process) show average values for these statistics when a series of length 6400 is split into 1, 2 or 4 separate series as described in text.

6 Non-identically distributed observations

There is a substantial literature on order statistics for random variables that are independent but not identically distributed (David and Nagaraja, 2003; Bon and Păltănea, 2006; Balakrishnan and Zhao, 2013), but it is not apparent how to directly use the results in these works to obtain a useful composite log-likelihood. The exact joint distribution of the order statistics for $X_1, \ldots, X_n$ independent but not identically distributed requires summing over all permutations of the indices $1, \ldots, n$ (David and Nagaraja, 2003)[Section 5.2]. However, even if this sum were feasible to compute, it would not generally be a good basis for inference because the order statistics are no longer a sufficient statistic when the observations are not identically distributed. This section

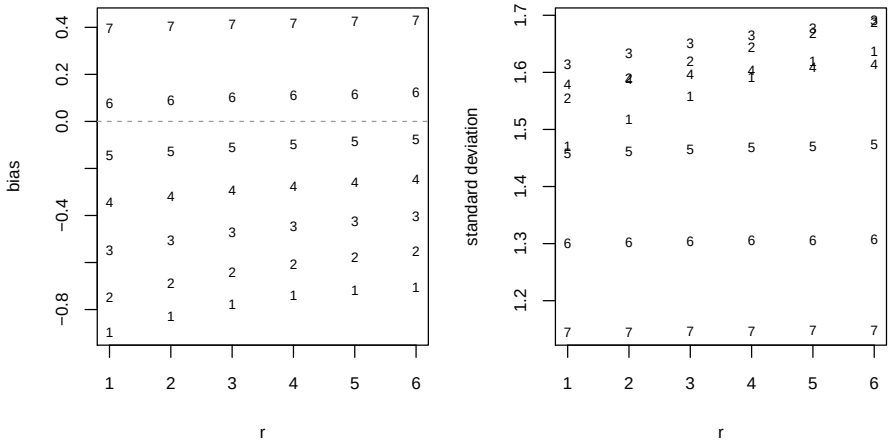

Fig. 6 Simulations of chi-squared process with Gamma(2,1) marginal distributions and first order autocorrelation 0.9. For series of length $n = 1500$, bias and standard deviations for estimates of $1 - 1/n$ quantile. Horizontal axis indicates r as defined in (26) and, writing ℓ for the plotting symbol, $j = 10 \times 2^{\ell-1}$ is the number of upper tail observations used to estimate the scale parameter.

suggests some possible WCLs that can be used for independent but not identically distributed observations that attempt to make acceptable compromises between computational feasiblity and statistical efficiency. One possibility is to use the conditional likelihood of the k'th order statistic given the values and indices of the first k order statistics. Let us assume no ties to avoid complications. Write $j(1), \ldots, j(n)$ for the indices of the ordered observations, so that, for example, $X_{j(1)} = X_{(1)}$ and define the set $J(k) = \{j(1), \ldots, j(k)\}$. Additionally, denote the density and survival functions for X_j by $f_{\theta,j}$ and $S_{\theta,j}$, respectively. Then the conditional density of $X_{(k)}$ given $X_{(1)}, \ldots, X_{(k-1)}$ and $J(k)$ is

$$\frac{f_{\theta,j(k)}(X_{(k)})}{S_{\theta,j(k)}(X_{(k-1)})} \cdot \prod_{i \notin J(k)} \frac{S_{\theta,i}(X_{(k)})}{S_{\theta,i}(X_{(k-1)})}. \tag{27}$$

Alternatively, at greater computational effort, we can calculate the conditional likelihood of $X_{(k)}$ given $X_{(1)}, \ldots, X_{(k-1)}$ and $J(k-1)$:

$$\begin{aligned} &\sum_{i \notin J(k-1)} \frac{f_{\theta,i}(X_{(k)})}{S_{\theta,i}(X_{(k-1)})} \cdot \prod_{\ell \notin \{J(k-1) \cup i\}} \frac{S_{\theta,\ell}(X_{(k)})}{S_{\theta,\ell}(X_{(k-1)})} \\ &= \sum_{i \notin J(k-1)} \frac{f_{\theta,i}(X_{(k)})}{S_{\theta,i}(X_{(k)})} \cdot \prod_{\ell \notin J(k-1)} \frac{S_{\theta,\ell}(X_{(k)})}{S_{\theta,\ell}(X_{(k-1)})}. \end{aligned} \tag{28}$$

Both of these formulas make sense for $k = 1$ as long as we define $X_{(0)} = -\infty$ and $J(0)$ as the null set. Weighted sums over k of the logarithms of either of (27) or (28) are possible WCLs. However, unlike the case for identically

distributed observations, neither of these give the exact log-likelihood when the weights are all 1, so there is the potential for loss of information when using such WCLs. Indeed, if the distributions of the observations are very different, then conditioning on $J(k)$ could lose a lot of information about θ. For example, suppose our model for the independent observations X_1 and X_2 is that X_i has a uniform distribution on $(\theta_i, \theta_i + 1)$, where (θ_1, θ_2) equals either $(0, 2)$ or $(2, 0)$. Then the actual observations X_1 and X_2 would tell us for sure which of the two possibilities for (θ_1, θ_2) was correct, whereas both (27) and (28) are identically 1 for $k = 1$ or 2 and thus provide no information about the parameters. Of course, this example is extreme but it makes the point that if the correct marginal distributions of the observations are very different, conditioning on the ranks of the observations can lead to a large loss of information.

One might hope that, because it conditions only on $J(k-1)$ rather than $J(k)$, WCLs based on (28) will sometimes yield more statistically efficient estimates than when using (27). Since both (27) and (28) reduce to the conditional density of $X_{(k)}$ given $X_{(1)}, \ldots, X_{(k-1)}$ when the observations are identically distributed, another approach to reduce this information loss is to preprocess the observations so that they are nearly identically distributed. Here is one possible way this preprocessing could be done. For some small $\delta > 0$, choose some threshold probability $1 - \delta$ and fit a quantile regression at this level to the observations using suitably chosen covariates that can explain the nonstationarity. Denote the fitted threshold for observation j by $\hat{t}_j$. Now fit a second quantile regression at level $1 - \delta/2$ using the same covariates as the previous regression. Write $\hat{r}_j$ for the difference in the fitted values of observation j for the levels $1 - \delta/2$ and $1 - \delta$ quantile regressions. Then write $Y_j = (X_j - \hat{t}_j)/\hat{r}_j$ for the normalized exceedances and fit a GPD with possibly varying scale and shape parameters using a weighted composite log-likelihood.

When the number of observations above a designated (possibly varying) threshold is large, calculating WCLs using even the simpler (27) could be onerous. We can reduce the number of computations substantially by breaking up the data into L groups and only considering the ranks within each group, similar to what was proposed in the previous section but for a different purpose. For example, in the analysis of daily precipitation data presented in the next section, I will split up the observations by months, which, in addition to reducing the computations by roughly a factor of 12, may help to reduce any information loss due to using order statistics of observations that are not identically distributed. More explicitly, define Γ_ℓ as the set of indices k that are in group ℓ and for which the normalized Y_k is positive. Write n_ℓ for the cardinality of this set. Let $Y_{(1,\ell)} < \cdots < Y_{(n_\ell,\ell)}$ be the order statistics among the Y_k's with $k \in \Gamma_\ell$ and define $j_\ell(k)$ and $J_\ell(k)$ as before except for just those observations in group ℓ. Then using, for example, (27), we have the WCL

$$\sum_{\ell=1}^{L} \sum_{k=1}^{n_\ell} \omega\left(\frac{k-1}{n_\ell}\right) \log\left\{ \frac{f_{\theta,j_\ell(k)}(Y_{(k,\ell)})}{S_{\theta,j_\ell(k)}(Y_{(k-1,\ell)})} \cdot \prod_{i \notin J_\ell(k)} \frac{S_{\theta,i}(Y_{(k,\ell)})}{S_{\theta,i}(Y_{(k-1,\ell)})} \right\}. \qquad (29)$$

7 Application to daily rainfall data

This section considers estimating seasonally varying upper quantiles of daily precipitation (measured to the nearest 0.01 inches) at the Central Park weather station in New York City for the period January 1, 1869–March 1, 2021, obtained from the National Centers for Environmental Information website (Menne et al., 2012). Figure 7 shows quantiles for these data by month. The months August–October have the smallest 0.75 quantiles and the largest 0.999 quantiles, so there is a clear seasonality in the data that cannot be explained by a simple scaling factor. Indeed, in October, the percentage of days with positive precipitation is 27.4 and the 0.75 quantile is 0.01 inch, the lowest possible positive value. It is worth noting that August–October is the peak of the North Atlantic hurricane season, although I have not checked to see which of the most extreme precipitation amounts are associated with hurricanes. Figure 8 suggests that there is some non-stationarity in the process across years, with the total annual precipitation taking on consistently low values from the late 1940s to the late 1960s and the last 50 years containing the 9 highest annual precipitation values out of the 152 complete years of data available. The maximum daily precipitation by year (right panel of Figure 8) does not show obvious non-stationarity and, to avoid complications, I will assume there is no long-term trend in upper quantiles across years. The temporal dependence in the data is fairly weak. Without correcting for seasonality, the raw lag one autocorrelation of the time series is 0.083 and the lag one autocorrelation of the indicator of positive precipitation is 0.161. Focusing on extreme precipitation, there are 30 days with 4 or more inches of rain in the dataset, only two of which occur on consecutive days, another pair of occurrences are four days apart and no other occurrences are within one month of each other. These modest dependencies should have a limited impact on estimates of marginal distributions and, given the simulation results for processes with much higher correlations in the previous section, are unlikely to induce substantial biases when using WCLs based on order statistics.

To provide a fair comparison between maximum likelihood estimates and maximum WCL estimates using (27), I split the data roughly in half into training and testing data in alternating periods of 4 years. Thus, the training data consists of 1869–1872, 1877–1880,..., 2013-2016 and all other years in the testing data, including the partial data for 2021. Spreading the years in the training data across the time period was done to reduce any influence of a changing climate on the results and four-year blocks were selected to reduce statistical dependence between the training and test data. In all quantile regressions and linear models for the logarithm of the scale parameter of the GPD, I used a periodic cubic spline basis with one-year period and 6 evenly spaced knots. For a range of quantiles $1-\delta$, using the `quantreg` package in R (Koenker, 2021), I then estimated threshold functions using quantile regression, yielding a fitted threshold function $\hat{t}_k$ as a function of day k. Because of the considerable length of this data set, I used the more accurate value of 365.242 days per year rather than 365.25, although it makes almost no differ-

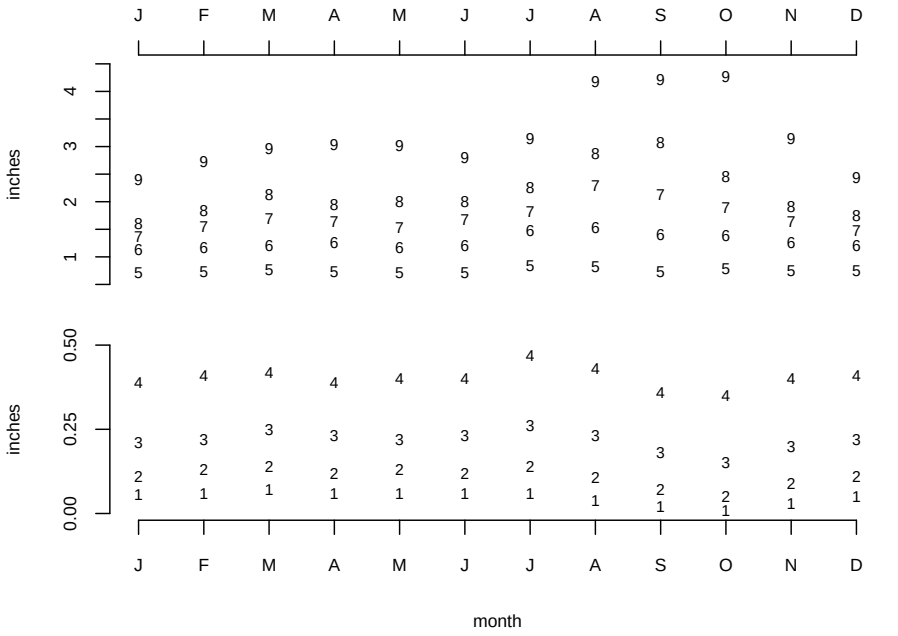

Fig. 7 Monthly quantiles of New York City daily precipitation. Symbols 1–9 correspond to 0.75, 0.8, 0.85, 0.9, 0.95, 0.98, 0.99, 0.995 and 0.999 quantiles, respectively. Vertical axis is split to allow easy visualization of monthly variations over full range of quantiles.

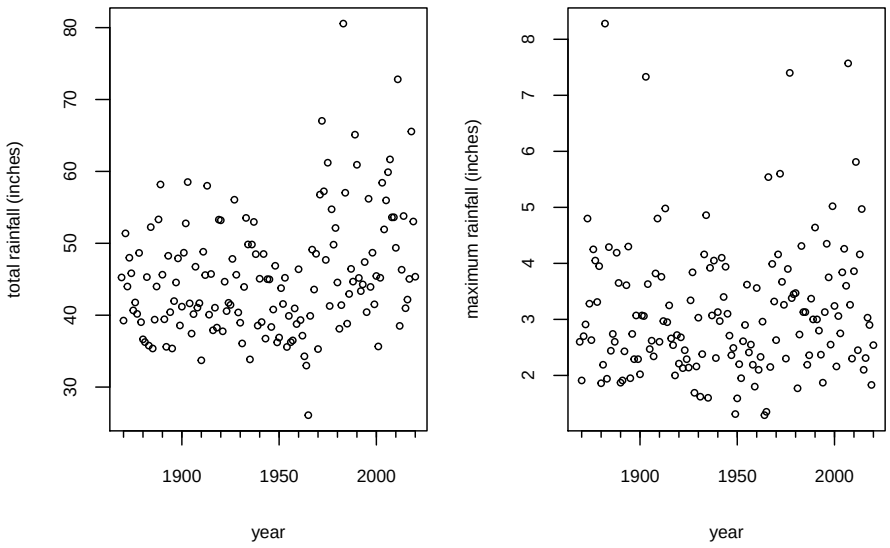

Fig. 8 By year, total precipitation and maximum daily precipitation in New York City.

ence in the results. For the ordinary (unweighted) likelihood, I then fit a GPD to the exceedances using a scale function $\hat{\sigma}_k$ whose logarithm is linear in the spline basis functions using the R package `extRemes` (Gilleland, 2020). For the WCLs only, to make the exceedances more nearly identically distributed, as described in the previous section, I normalized the exceedances by dividing by the difference in the fitted values for each day for quantile regressions at quantile $1-\delta/2$ and $1-\delta$. Denote this normalizing function by $\hat{r}_k$. Using (29), I then fit GPDs to the normalized exceedances over these seasonally varying thresholds again using a linear model for the logarithm of the scale function $\hat{\sigma}_k$. Optimization was carried out using the `nlm` command in R and starting values given by the ordinary likelihood estimates for the given value of δ. I considered allowing the shape parameter to vary seasonally by allowing a separate value of this parameter for the months of August–October, but preliminary analyses indicated including such a factor did little to improve the fit, so I chose to use a constant shape parameter.

Although the distribution of extreme precipitation clearly varies seasonally, it may be more relevant to assess the chances of precipitation events of a given size regardless of season. Thus, to evaluate these fits for these various fitted models, I considered how well these models fit precipitation amounts above either 2 or 3 inches. Precipitation above 2 inches occurs 756 times (0.63% of days) and precipitation greater than 3 inches 97 times (0.17% of days) over the entire dataset.

Following Stein (2021), the criterion function I used is the log-likelihood of the testing data censored from below at a cutoff c of either 2 or 3 inches, ignoring any possible dependence in the observations. Now using k to indicate a day in the testing set and writing p_k for the corresponding observed precipitation, this criterion function can be written as

$$\sum_{p_k \leqslant c} \log \left[1 - \delta \left(1 + \frac{\hat{\xi}}{\hat{\sigma}_k \hat{r}_k} (c - \hat{t}_k) \right)^{-1/\hat{\xi}} \right] + \sum_{p_k > c} \left[\log \frac{\delta}{\hat{\sigma}_j \hat{r}_k} - \left(\frac{1}{\hat{\xi}} - 1 \right) \log \left(1 + \frac{\hat{\xi}}{\hat{\sigma}_k \hat{r}_k} (p_k - \hat{t}_k) \right) \right], \tag{30}$$

where $\hat{r}_k$ is identically 1 for the maximum likelihood estimate. This result assumes that $\hat{t}_k < c$ for all j, which is true in all cases shown here. Note that $\hat{\sigma}_k \hat{r}_k$ is the estimated scale parameter of the GPD on day k for the unnormalized observations.

These criterion functions are plotted in Figure 9. For the 2 inch cutoff, the best maximum likelihood estimate beats the best WCL estimate by about 1 log-likelihood unit. However, the performance of the maximum likelihood estimate varies quite irregularly in δ, so this modestly superior performance could easily be due to luck. In practice, it may be desirable to have a procedure whose results are fairly insensitive to modest changes in the difficult to specify quantity δ and, in this regard, the weighted procedure clearly dominates. When the more extreme and, thus, perhaps more relevant cutoff of 3 inches is used,

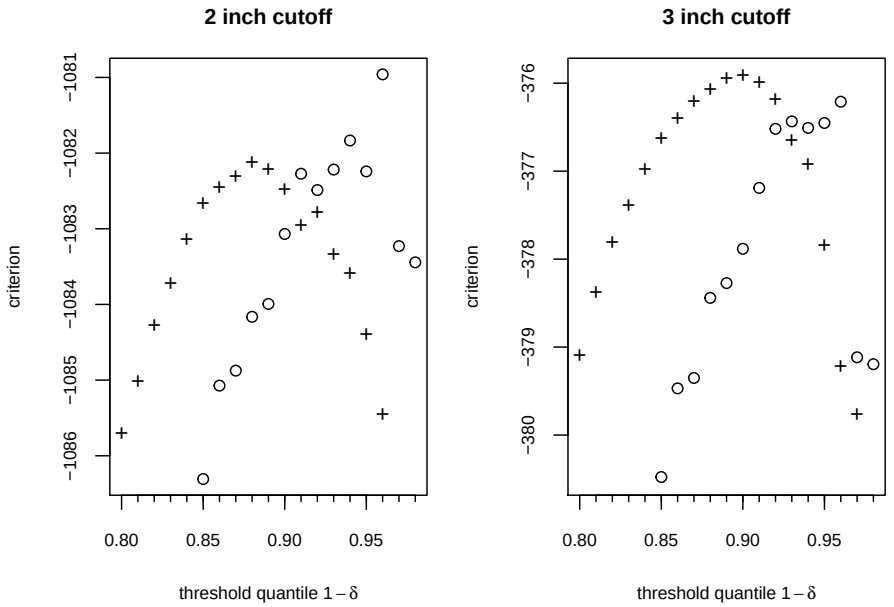

Fig. 9 Cross-validated log-likelihoods for New York City daily precipitation censored at 2 and 3 inches based on a seasonally varying threshold fit via quantile regression at a range of quantiles and a seasonally varying GPD for exceedances over thresholds using maximum likelihood (○) and maximum WCL (+) with a linear weight function.

the best WCL estimate is marginally better than the best ordinary likelihood estimate. Furthermore, with the 3 inch cutoff, the ordinary likelihood estimate performs substantially worse when $1 - \delta$ is increased from its optimal value, 0.96, to the next larger value considered of 0.97.

Figure 10 gives more detailed results for the number, by month, of observed and expected exceedances of 2 and 3 inches of daily precipitation. The year 2021 is left out of the calculations so that the training and testing data have the same number of days, although, in fact, there were no observed exceedances of 2 inches for the partial year 2021. Because I took a year to have 365.242 days, the expected number of exceedances in the testing and training periods are not exactly equal, but the difference is trivial and is ignored in this figure. Based on their good performance in Figure 9, this figure shows results for $\delta = 0.04$ for the unweighted procedure and $\delta = 0.11$ for the linear weighted procedure. For both cutoffs, the expected seasonal patterns are very similar for the two estimates, with differences that are much smaller than the differences between the observed outcomes in the training and testing periods. If we go to even higher cutoffs, the observed exceedances are too rare to make a monthly comparison informative. Summing over all months, the unweighted and weighted procedures estimate the expected number of exceedances of 4 inches over the training (or testing) period of 15.16 and 15.54, respectively,

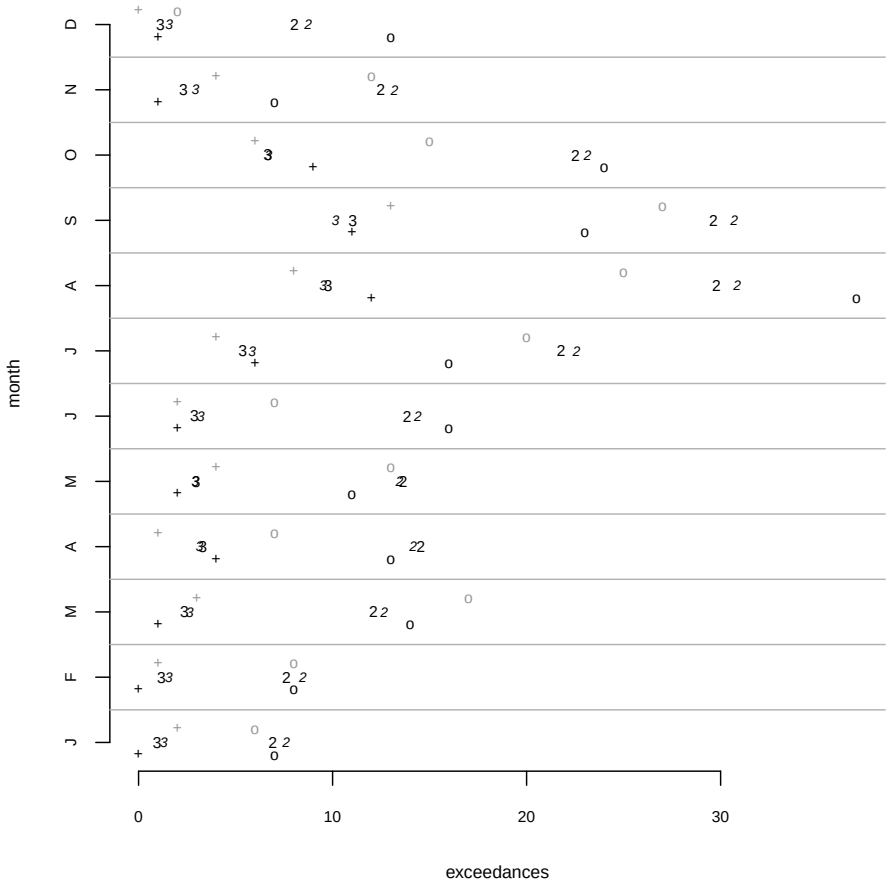

Fig. 10 Expected and observed number of exceedances by month of daily precipitation greater than 2 and greater than 3 inches. Observed exceedances given by "o" for 2 inches and "+" for 3 inches; black symbols are for the training period and gray symbols for the testing period. Expected exceedances use the exceedance cutoff as the plotting symbol, where the larger roman font corresponds to the ordinary maximum likelihood estimate with $\delta = 0.04$ and the smaller italicized font to the WCL with linear weight function and $\delta = 0.11$.

whereas the observed number of exceedances was 18 for the training period and 12 for the testing period. The same results for expected exceedances of 5 inches are 5.15 and 5.27 for the unweighted and weighted procedures, respectively; there are 4 exceedances of 5 inches in both the training and testing periods. Overall, both procedures give very good agreement with the observed record in both the training and testing periods.

This kind of cross-validation could be used to select δ, although there are many issues that would need to be addressed before such an approach could be used routinely in practice. One simple approach to choosing the threshold quantile $1 - \delta$ when ξ is constant is to plot estimates of ξ as a function of this

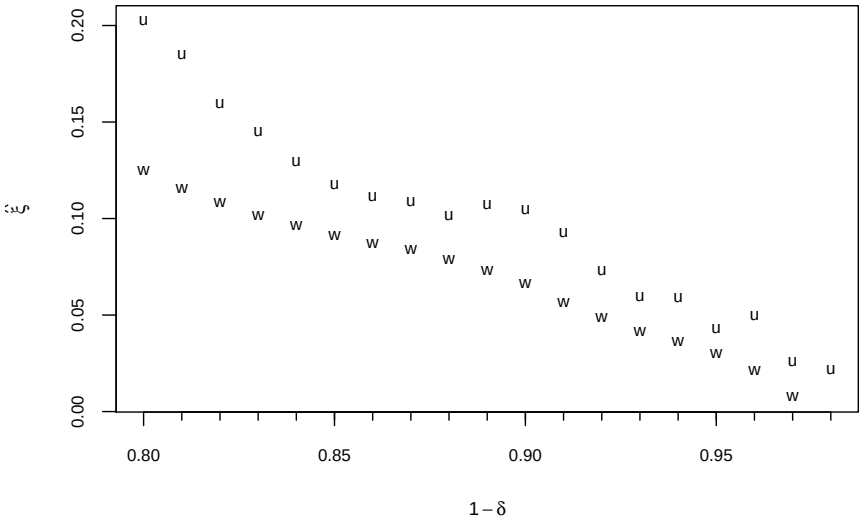

Fig. 11 Estimated shape parameters as a function of the threshold quantile $1-\delta$ for the ordinary log-likelihood (u) and WCL with linear weight function (w). Value for WCL and $1-\delta = 0.98$ not shown because of convergence issues in fitting model.

quantile and then select δ at a point where the estimated shape parameter shows no obvious systematic variation for smaller δ (Coles, 2001)[p. 83]. Figure 11 gives such a plot for the two procedures. For both procedures, it appears that the estimated shape is decreasing across the entire range of δ values, which would normally indicate that δ should be taken even smaller than 0.02, the minimum value considered. However, the cross-validation results indicate that such small values of δ are not the best for fitting the upper tail of the rainfall distribution. We do see that the estimates of shape using the WCL show less local variation than the ordinary log-likelihood. For example, while the overall slope of the estimates for ordinary log-likelihood is steeper than for the WCL, the WCL estimates are monotonically decreasing in $1-\delta$, whereas the estimates based on the ordinary log-likelihood are not. Thus, it may be easier to spot systematic trends in shape estimates as the threshold varies when using WCL, although this example indicates that, similar to what was found in Figure 4, one might be better off deliberately selecting a threshold at which the estimated shape is still clearly systematically changing.

8 Discussion

The various WCLs described here do not require assuming a GPD model for the upper tail of the distribution. There are a number of approaches to modeling distribution functions of iid observations when the main interest is in the upper tail, including mixture modeling (Scarrott and MacDonald, 2012; Scarrott, 2016), parametric extensions of the GPD (Papastathopoulos and Tawn,

2013; Naveau et al., 2016; Stein, 2020, 2021) and semiparametric approaches (Huang et al., 2018). In this setting, one would presumably consider choosing j in (3) much larger than when fitting a GPD, perhaps even taking $j = n$ but still taking w_k small when k is a substantial fraction of n to limit the influence of the smaller observations. For temperature, one may sometimes be interested in both the upper and lower tails of the distributions, in which case the models in Stein (2020, 2021) may be appropriate. In this case, taking n odd for convenience, one might consider a WCL that uses the marginal log-likelihood of the median and then includes two sums like the sum in (3), one for the order statistics increasing from the median and one for those decreasing from the median. Further afield, this WCL could be used to downweight the more extreme rather than the less extreme observations, giving estimates of characteristics of distributions that are robust to outliers. It could be interesting to compare the resulting estimates to traditional robust procedures (Huber, 2004).

Clearly more can be done to extend the theory and to refine the methods proposed here for practical use. Some specific issues needing attention include the preprocessing step when observations are not identically distributed and methods for selecting thresholds. Nevertheless, the results obtained here show that WCLs based on order statistics can have some noticeable, if modest, benefits over ordinary log-likelihoods for fitting GPDs to tail observations, both for parameter and extreme quantile estimation. The weighted approach can substantially reduce random fluctuations in estimated parameters and extreme quantiles as a function of the threshold. This reduction in fluctuations might be expected to yield improved extreme quantile estimation when thresholds are chosen in a data-driven manner, but since threshold selection is often done based on visual inspection of graphical diagnostics, it is not clear how one might definitively reach such a finding.

9 Proof of Theorem 1

Writing $\lfloor \cdot \rfloor$ for the greatest integer function and setting c_n in Theorem 2.1 of Drees (1998) to $1/(\alpha\sqrt{j_n})$, it follows that there exists a sequence of Brownian motions $\{B_n(t)\}_{t \geqslant 0}$ such that for each $\epsilon > 0$,

$$\sup_{0<t\leqslant 1} t^{1/2+\epsilon} \left| \alpha\sqrt{j_n} \left(X_{(n-\lfloor j_n t \rfloor)} - Q\left(1 - \frac{j_n}{n}\right) + \frac{1}{\alpha} \log t \right) + t^{-1} B_n(t) - \sqrt{j_n} \Phi\left(\frac{j_n}{n}\right) \Psi(t) \right| \to 0 \tag{31}$$

in probability as $n \to \infty$. From the assumptions on ω, we have

$$j_n^{-1} W_{j_n} = 1 + O\big(j_n^{-1}\big). \tag{32}$$

Writing Δ_k for $kw_k - (k-1)w_{k-1}$, summation by parts implies

$$\begin{aligned}\sum_{k=1}^{j_n} \Delta_k \log \frac{k}{j_n} &= \sum_{k=1}^{j_n-1} k w_k \log \frac{k}{k+1} \\ &= -\sum_{k=1}^{j_n-1} k w_k \left(\frac{1}{k} + O\left(\frac{1}{k^2}\right)\right) = -W_{j_n} + O(\log j_n). \end{aligned} \tag{33}$$

Define $\nu(t) = \omega(t) + t\omega'(t)$, so that $\Delta_k = \nu(k/j_n) + O(1/j)$, uniformly in k. Using (31) and(32), it follows that

$$\begin{aligned} & W_{j_n}^{-1} \sum_{k=1}^{j_n} \left(\Delta_k - \nu\left(\frac{k}{j_n}\right)\right)\left(X_{(n-k+1)} - Q\left(1 - \frac{j_n}{n}\right) + \frac{1}{\alpha}\log\frac{k}{j_n}\right) \\ &= O_p\left(j^{-2}\sum_{k=1}^{j}\left|X_{(n-k+1)} - Q\left(1-\frac{j_n}{n}\right) + \frac{1}{\alpha}\log\frac{k}{j_n}\right|\right) \\ &= O_p\left(j_n^{-5/2}\sum_{k=1}^{j_n}\left|-\frac{j_n}{k}B_n\left(\frac{k}{j_n}\right) + \sqrt{j_n}\Phi\left(\frac{j_n}{n}\right)\Psi\left(\frac{k}{j_n}\right)\right|\right) \\ &= O_p\left(j_n^{-5/2}\sum_{k=1}^{j_n}\left(\frac{j_n}{k}\right)^{1/2}\right) \\ &= O_p\big(j_n^{-3/2}\big). \end{aligned} \tag{34}$$

It follows from (6), (33) and (34) that

$$\begin{aligned} & W_{j_n}^{-1}\sum_{k=1}^{j_n}\Delta_k\left(X_{(n-k+1)} - Q\left(1-\frac{j_n}{n}\right) + \frac{1}{\alpha}\log\frac{k}{j_n}\right) \\ &= W_{j_n}^{-1}\sum_{k=1}^{j_n}\nu\left(\frac{k}{j_n}\right)\left(X_{(n-k+1)} - Q\left(1-\frac{j_n}{n}\right) + \frac{1}{\alpha}\log\frac{k}{j_n}\right) + O_p\big(j_n^{-3/2}\big). \end{aligned} \tag{35}$$

Similar calculations show that

$$\begin{aligned} & W_{j_n}^{-1}\sum_{k=1}^{j_n}\nu\left(\frac{k}{j_n}\right)\left(X_{(n-k+1)} - Q\left(1-\frac{j_n}{n}\right) + \frac{1}{\alpha}\log\frac{k}{j_n}\right) \\ &= j_n W_{j_n}^{-1}\int_0^1 \nu(t)\left(X_{(n-\lfloor tj_n\rfloor)} - Q\left(1-\frac{j_n}{n}\right) + \frac{1}{\alpha}\log t\right)dt + O_p\left(\frac{\log j_n}{j_n}\right). \end{aligned} \tag{36}$$

Setting $\epsilon = 1/4$ in (31),

$$\begin{aligned} & \int_0^1 \nu(t)\left(X_{(n-\lfloor tj\rfloor)} - Q\left(1-\frac{j_n}{n}\right) + \frac{1}{\alpha}\log t\right)dt \\ &= \int_0^1 \frac{\nu(t)}{t^{3/4}}\frac{t^{3/4}}{\alpha\sqrt{j}}\left(t^{-1}B_n(t) - \sqrt{j}\Phi\left(\frac{j_n}{n}\right)\Psi(t)\right)dt + o_p(j_n^{-1/2}) \end{aligned} \tag{37}$$

since $\nu(t)t^{-3/4}$ is integrable on $(0,1)$.

Putting together (12), (32) and (34)–(37),

$$W_{j_n}^{-1}\sum_{k=1}^{j_n}\Delta_k\left(X_{(n-k+1)}-Q\left(1-\frac{j_n}{n}\right)+\frac{1}{\alpha}\log\frac{k}{j_n}\right)$$
$$=\frac{1}{\alpha\sqrt{j_n}}\int_0^1\nu(t)\left(t^{-1}B_n(t)-C\Psi(t)\right)\,dt+o_p(j_n^{-1/2}). \tag{38}$$

From (31) with $t=1$, we have $\alpha\sqrt{j}(X_{(n-j_n)}-Q(1-j_n/n))-B_n(1)$ converges to 0 in probability, which, together with (32), (33), (38) and the formula for $\hat{Q}$ in (7), implies

$$\begin{aligned}
&\hat{Q}(1-\delta_n)\\
&=\left(X_{(n-j_n)}-Q\left(1-\frac{j_n}{n}\right)+Q\left(1-\frac{j_n}{n}\right)\right)\left(1+\frac{j\omega(1)}{W_{j_n}}\log\frac{\delta_n(n+1)}{j_n+1}\right)\\
&\quad-W_{j_n}^{-1}\log\frac{\delta_n(n+1)}{j_n+1}\sum_{k=1}^{j_n}\Delta_k\left(X_{(n-k+1)}-Q\left(1-\frac{j_n}{n}\right)+\frac{1}{\alpha}\log\frac{k}{j_n}\right)\\
&\quad-W_{j_n}^{-1}\log\frac{\delta_n}{j_n+1}\sum_{k=1}^{j}\Delta_k\left(Q\left(1-\frac{j_n}{n}\right)-\frac{1}{\alpha}\log\frac{k}{j_n}\right)\\
&=\log\frac{\delta_n}{j_n+1}\left[\frac{\omega(1)}{\alpha\sqrt{j_n}}B_n(1)-\frac{1}{\alpha\sqrt{j_n}}\int_0^1\frac{\nu(t)}{t}B_n(t)\,dt\right.\\
&\qquad\left.+\frac{C}{\alpha\sqrt{j_n}}\int_0^1\nu(t)\Psi(t)\,dt-\frac{1}{\alpha}+o_p\big(j_n^{-1/2}\big)\right]+Q\left(1-\frac{j_n}{n}\right).
\end{aligned}\tag{39}$$

Integration by parts yields

$$\int_0^1\nu(t)\Psi(t)\,dt=-\gamma\int_0^1 t^{\gamma}\omega(t)\,dt. \tag{40}$$

From (12), we have

$$j_n\sim\left(\frac{Ca_1^{\gamma+1}n^{\gamma}}{a_2}\right)^{1/(\gamma+1/2)}, \tag{41}$$

which, together with (10), implies

$$\sqrt{j_n}\left(Q\left(1-\frac{j_n}{n}\right)-\frac{1}{\alpha}\log\frac{n\delta_n}{j_n}-Q(1-\delta_n)\right)\to\frac{C}{\alpha}. \tag{42}$$

To prove the theorem, we only need that this limit is finite, but we will want the sharper result to obtain a heuristic correction term to (13). For $B(\cdot)$ Brownian motion, we then have

$$\begin{aligned}
&\frac{\sqrt{j_n}}{\log(j_n/(n\delta_n))}\left(\hat{Q}_n(1-\delta_n)-Q(1-\delta_n)\right)\\
&\qquad\to-\frac{\omega(1)}{\alpha}B(1)+\frac{1}{\alpha}\int_0^1\frac{\nu(t)}{t}B(t)\,dt-\frac{C}{\alpha}\int_0^1 t^{\gamma}\omega(t)\,dt
\end{aligned}\tag{43}$$

in distribution. Theorem 1 follows from

$$\begin{aligned}
&\mathrm{Var}\left(-\omega(1)B(1) + \int_0^1 \frac{\nu(t)}{t} B(t)\,dt\right) \\
&\quad = \omega(1)^2 - 2\omega(1)\int_0^1 \nu(t)\,dt + \int_0^1\int_0^1 \frac{\nu(t)\nu(s)}{ts}\min(s,t)\,ds\,dt \\
&\quad = -\omega(1)^2 + 2\int_0^1 \frac{\nu(t)}{t}\left[\int_0^t \nu(s)\,ds\right] dt \\
&\quad = -\omega(1)^2 + \int_0^1 \left[\frac{d}{dt}\left(t\omega(t)^2\right) + \omega(t)^2\right] dt \\
&\quad = \int_0^1 \omega(t)^2\,dt. \qquad \square
\end{aligned}$$

Acknowledgements This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) under Contract DE-AC02-06CH11347.

Data availability statement Data used in this study is freely available from the National Centers for Environmental Information at the website https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail.

References

Anderes EB, Stein ML (2011) Local likelihood estimation for nonstationary random fields. Journal of Multivariate Analysis 102(3):506–520, DOI https://doi.org/10.1016/j.jmva.2010.10.010

Balakrishnan N, Zhao P (2013) Ordering properties of order statistics from heterogeneous populations: a review with an emphasis on some recent developments. Probability in the Engineering and Informational Sciences 27(4):403–443, DOI 10.1017/S0269964813000156

Beirlant J, Caeiro F, Gomes M (2012) An overview and open research topics in statistics of univariate extremes. REVSTAT 10:1–31

Bon JL, Păltănea E (2006) Comparison of order statistics in a random sequence to the same statistics with i.i.d. variables. ESAIM: PS 10:1–10, DOI 10.1051/ps:2005020

Caeiro F, Henriques-Rodrigues L, Prata Gomes D (2019) A simple class of reduced bias kernel estimators of extreme value parameters. Computational and Mathematical Methods 1(3):e1025, DOI https://doi.org/10.1002/cmm4.1025

Chavez-Demoulin V, Davison AC (2012) Modelling time series extremes. REVSTAT 10:109–133

Coles S (2001) An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, London

Csorgo S, Deheuvels P, Mason D (1985) Kernel estimates of the tail index of a distribution. The Annals of Statistics 13(3):1050–1077, URL http://www.jstor.org/stable/2241125

David HA, Nagaraja HN (2003) Order Statistics, 3rd edn. Wiley-Interscience, Hoboken, NJ

Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternatives. Biometrika 74(1):33–43, URL http://www.jstor.org/stable/2336019

Davison A, Huser R, Thibaud E (2013) Geostatistics of dependent and asymptotically independent extremes. Mathematical Geosciences 45:511–529, URL https://doi.org/10.1007/s11004-013-9469-y

Davison AC, Smith RL (1990) Models for exceedances over high thresholds. Journal of the Royal Statistical Society: Series B (Methodological) 52(3):393–425, DOI 10.1111/j.2517-6161.1990.tb01796.x

Devroye L, Györfi L (1985) Nonparametric Density Estimation: The L_1 View. John Wiley & Sons, New York

Drees H (1998) On smooth statistical tail functionals. Scandinavian Journal of Statistics 25(1):187–210, DOI https://doi.org/10.1111/1467-9469.00097

Epanechnikov VA (1969) Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications 14(1):153–158, URL https://doi.org/10.1137/1114019

Falk M (1985) Asymptotic Normality of the Kernel Quantile Estimator. The Annals of Statistics 13(1):428–433, DOI 10.1214/aos/1176346605

Fawcett L, Walshaw D (2007) Improved estimation for temporally clustered extremes. Environmetrics 18(2):173–188, DOI https://doi.org/10.1002/env.810

Fawcett L, Walshaw D (2012) Estimating return levels from serially dependent extremes. Environmetrics 23(3):272–283, DOI https://doi.org/10.1002/env.2133

Ferro CAT, Segers J (2003) Inference for clusters of extreme values. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65(2):545–556, DOI https://doi.org/10.1111/1467-9868.00401

Gilleland E (2020) R package extremes. URL https://cran.r-project.org/web/packages/extRemes/

Gomes DP, Neves MM (2015) Bootstrap and other resampling methodologies in statistics of extremes. Communications in Statistics - Simulation and Computation 44(10):2592–2607, DOI 10.1080/03610918.2014.895834

Groeneboom P, Lopuhaä H, de Wolf P (2003) Kernel-type estimators for the extreme value index. The Annals of Statistics 31(6):1956–1995, DOI 10.1214/aos/1074290333

de Haan L, Ferreira A (2006) Extreme Value Theory: An Introduction. Springer, New York

Huang WK, Nychka DW, Zhang H (2018) Estimating precipitation extremes using the log-histospline. Environmetrics 0(0):e2543, DOI 10.1002/env.2543, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2543

Huber PJ (2004) Robust Statistics. John Wiley & Sons, Hoboken, NJ

Jones MC, Signorini DF (1997) A comparison of higher-order bias kernel density estimators. Journal of the American Statistical Association 92(439):1063–1073, DOI 10.1080/01621459.1997.10474062

Koenker R (2021) R package quantreg. URL https://cran.r-project.org/web/packages/quantreg

Lawrance AJ (1990) Discussion of the paper by Davison and Smith. Journal of the Royal Statistical Society: Series B (Methodological) 52(3):436–437, DOI https://doi.org/10.1111/j.2517-6161.1990.tb01797.x

Lawrance AJ, Lewis PAW (1981) A new autoregressive time series model in exponential variables (NEAR(1)). Advances in Applied Probability 13(4):826–845, URL http://www.jstor.org/stable/1426975

Li B, Babu G (2019) A Graduate Course in Statistical Inference. Springer, New York, NY

Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG (2012) An overview of the global historical climatology network-daily database. Journal of Atmospheric and Oceanic Technology 29:897–910, DOI doi:10.1175/JTECH-D-11-00103.1

Naveau P, Huser R, Ribereau P, Hannart A (2016) Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. Water Resources Research 52(4):2753–2769, DOI https://doi.org/10.1002/2015WR018552

Olver F, Lozier D, Boisvert R, Clark C (2010) The NIST Handbook of Mathematical Functions. Cambridge University Press, New York, NY

Papastathopoulos I, Tawn JA (2013) Extended generalised Pareto models for tail estimation. Journal of Statistical Planning and Inference 143(1):131–143, DOI https://doi.org/10.1016/j.jspi.2012.07.001, URL http://www.sciencedirect.com/science/article/pii/S0378375812002388

Rao TS (1970) The fitting of non-stationary time-series models with time-dependent parameters. Journal of the Royal Statistical Society: Series B (Methodological) 32(2):312–322, DOI https://doi.org/10.1111/j.2517-6161.1970.tb00844.x

Rényi A (1953) On the theory of order statistics. Acta Mathematica Academiae Scientiarum Hungaricase 4:191–231

Rosenblatt M (1971) Curve estimates. Biometrika 42(6):1815–1842

Scarrott C (2016) Univariate extreme value mixture modeling. In: Dey D, Yan J (eds) Extreme Value Modeling and Risk Analysis: Methods and Applications, CRC Press, Boca Raton, FL, chap 3, pp 41–67

Scarrott C, MacDonald A (2012) A review of extreme value threshold estimation and uncertainty quantification. REVSTAT 10:33–60

Sheather SJ (2004) Density estimation. Statistical Science 19(4):588–597, URL http://www.jstor.org/stable/4144429

Stein ML (2020) Parametric models for distributions when interest is in extremes with an application to daily temperature. Extremes URL https://doi.org/10.1007/s10687-020-00378-z

Stein ML (2021) A parametric model for distributions with flexible behavior in both tails. Environmetrics 32(2):e2658, DOI https://doi.org/10.1002/env.

2658

Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Statistica Sinica 21(1):5–42, URL http://www.jstor.org/stable/24309261