The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1524–1549
https://doi.org/10.1214/21-AOAS1555

IMPROVED INFERENCE ON RISK MEASURES FOR UNIVARIATE EXTREMES

BY LÉO R. BELZILE[1,a] AND ANTHONY C. DAVISON[2,b]

[1]*Department of Decision sciences, HEC Montréal,* [a]*leo.belzile@hec.ca*
[2]*Institute of Mathematics, École polytechnique fédérale de Lausanne,* [b]*anthony.davison@epfl.ch*

We discuss the use of likelihood asymptotics for inference on risk measures in univariate extreme value problems, focusing on estimation of high quantiles and similar summaries of risk for uncertainty quantification. We study whether higher-order approximation, based on the tangent exponential model, can provide improved inferences. We conclude that inference based on maxima is generally robust to mild model misspecification and that profile likelihood-based confidence intervals will often be adequate, whereas inferences based on threshold exceedances can be badly biased but may be improved by higher-order methods, at least for moderate sample sizes. We use the methods to shed light on catastrophic rainfall in Venezuela, flooding in Venice, and the lifetimes of Italian semisupercentenarians.

1. Introduction.

1.1. *Risk measures*. Estimating worst-case scenarios is important for risk management and policy making, but the hypothetical events that keep decision-makers awake at night can lie far outside the available data. Large-sample likelihood approximations are routinely used for inference based on extreme observations, but the numbers of rare events can be small, which raises the question of the adequacy of standard asymptotic approximations. Improved approximations are used in other domains (e.g., Brazzale, Davison and Reid (2007)), but thus far they have had limited impact in extreme-value statistics.

The focus in applications of extremes is typically on measures of risk, such as exceedance probabilities, quantiles, or other summaries of the distribution tail. Accurate small-sample inference for such quantities, which we denote by ψ in general discussion, is the topic of this paper. Their estimators may have very asymmetric sampling distributions, so classical "estimate $\pm c\times$ standard error" confidence intervals may be appallingly bad: they may contain inadmissible parameter values, and the empirical probability that such an interval contains the true parameter value may be much less than the nominal probability. One approach to dealing with these deficiencies is to compute the confidence interval on a transformed scale, for example, by considering the logit of a probability, but even this can perform very badly, as we shall see below. In such settings it is natural to require that confidence intervals are invariant to so-called "interest-preserving transformations" which transform in a natural way: if (L, U) is a $(1-\alpha)$ confidence interval for a scalar parameter ψ in a model with other parameters λ, then, for any monotone increasing transformation g, $[g(L), g(U)]$ is the corresponding confidence interval for $g(\psi)$, even if the remaining parameters are transformed from λ to $\zeta(\lambda, \psi)$. Confidence intervals based on the profile likelihood have this property, and this is an important reason to favor these intervals in applications, but we shall see below that modified versions of them may be preferred in some cases.

Received January 2021; revised July 2021.
Key words and phrases. Extreme value distribution, generalized Pareto distribution, higher-order asymptotic inference, Poisson process, tangent exponential model, profile likelihood.

1.2. *Motivating example*: *Vargas tragedy.* Cumulative rainfall of around 911 mm over a three-day period in mid-December 1999 led to landslides and debris flow that caused an estimated 30,000 deaths in the Venezuelean coastal state of Vargas. Daily cumulated rainfall data recorded at the Maiquetía *Simón Bolívar International Airport* for the years 1961–1999 were analyzed in Coles and Pericchi (2003) and Coles, Pericchi and Sisson (2003), whose fit to annual maxima up to 1998 suggested that the return period for such an event would be approximately 18 million years, though more sophisticated models led to much more reasonable risk estimates. Yearly maxima for 1951–1960 and anecdotal records are also available: for example, during the floods of February 1951, a reported 282 mm of rain fell in Maiquetía over consecutive days, while the neighbouring station of El Infiernito in the Cordillera de la Costa, between Caracas and Maiquetía, recorded 529 mm for the same day (Wieczorek et al. (2001)). These events suggest potential for extremes well beyond the range seen in the daily records.

To motivate the practical need for modified likelihood approximations, we fit the generalized Pareto and the inhomogeneous Poisson process models described, respectively, in Sections 2.1 and 4.1 to daily rainfall totals from 1961 to November 1999 that exceed 27 mm and take the risk measure ψ to be the median of the semicentennial maximum distribution. The threshold stability plot in the right-hand panel of Figure 1 suggests that a threshold of $u = 27$ mm is appropriate, and the extremogram (Davis and Mikosch (2009)), which estimates the conditional probabilities $\mathsf{P}(Y_t > u \mid Y_{t-h} > u)$ for a threshold u and lags h, suggests that high rainfall on successive days is only very weakly dependent. This leaves $n_u = 142$ exceedances for inference, on average 3.74 per year. The profile likelihoods for ψ, shown in Figure 1, are nearly indistinguishable and highly asymmetric; the maximum likelihood estimator is $\hat{\psi} = 153$ mm, the 95% profile confidence intervals are (116,262) mm, and higher-order confidence intervals are (118,278) mm (generalized Pareto) and (125,301) mm (Poisson process); the latter accounts for the randomness of the number of exceedances. In view of the shape of the log-likelihood, any symmetric interval would be highly inappropriate and could lead to severe underestimation of the risk of rare events. The higher-order intervals are both more asymmetric and shifted to the right relative to the profile likelihood intervals, corresponding

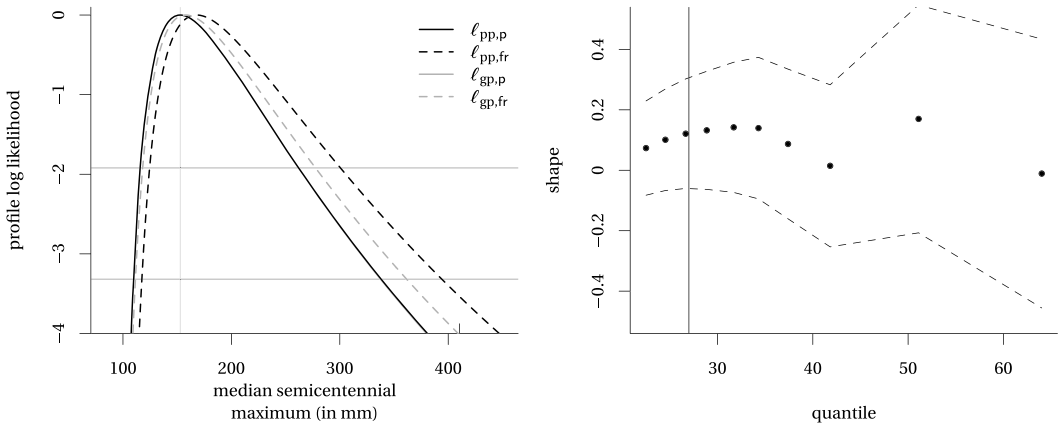

FIG. 1. *Left panel*: *Profile likelihoods* (*solid*) *and higher-order versions* (*dashed*), *discussed in Section* 3.2, *for the median semicentennial maximum daily rainfall at Maiquetía based on threshold exceedances of daily cumulated rainfall above* 27 *mm for* 1961–1999 (*excluding December* 1999) *using the generalized Pareto* (*grey*) *and inhomogeneous Poisson process* (*black*) *likelihoods. The dashed grey horizontal line at* -1.92 (-3.32) *indicates cutoff values for* 95% (99%) *confidence intervals based on the asymptotic* χ^2_1 *distribution. Right panel*: *Threshold stability plot of Wadsworth* (2016) *for the shape parameter* ξ *of the generalized Pareto distribution with* 95% *simultaneous confidence intervals.*

to substantially increased risk. Simulations summarized in Section 5 suggest that the modified profile likelihood interval has better properties than the usual profile likelihood interval, both in this case and for related measures of risk.

1.3. *Outline.* In later sections we first outline the necessary elements of extreme-value statistics and their use for the estimation of risk. Then, we explain the construction of variants of the profile likelihood using the tangent exponential model approximation, and show how they influence the conclusions to be drawn in three applied settings: rainfall extremes in coastal Venezuela, probabilities of severe flooding in Venice, and excess lifetimes of Italian semisupercentenarians. Finally we use simulation to assess when the higher-order methods provide improved inferences. We conclude that the coverage properties of profile likelihood intervals are adequate overall, though small-sample bias appears for extrapolation too far into the tail. The improved methods yield wider confidence intervals with more accurate error rates, though slightly larger samples are needed for them to be effective. The code used for the simulation study and the applications is available for download at https://github.com/lbelzile/hoa-extremes

2. Basic notions.

2.1. *Extremal models.* Extreme value analysis is concerned with two main problems: estimating the probability of extremes of given sizes and estimating a typical worst-case scenario over a given period. A standard approach to addressing them is to fit specific distributions justified by asymptotic arguments to maxima (or minima) over specific time periods or to exceedances of a high (or low) threshold. The limiting distributions for high and low extremes are related by a simple change of sign, so we can consider only maxima and exceedances of a high threshold.

The extremal types theorem (Fisher and Tippett (1928); Gnedenko (1943)) characterizes the limiting distribution of maxima under very mild conditions, but we use slightly stronger assumptions for ease of exposition. Let $F(y)$ denote a thrice-differentiable distribution function with density $f(y)$ whose support has upper endpoint y^*, define $s(y) = -F(y)\log\{F(y)\}/f(y)$, let b_m denote the solution of the equation $-\log F(b_m) = m^{-1}$, and let $a_m = s(b_m) > 0$. If M_m denotes the maximum of a block of m independent observations from F, then the existence of $\xi^\star = \lim_{m\to\infty} s'(b_m)$ implies the existence of the limit

$$\lim_{m\to\infty} \mathsf{P}\{(M_m - b_m)/a_m \le y\} = \lim_{m\to\infty} F^m(a_m x + b_m) = \exp\{-(1+\xi^\star y)_+^{-1/\xi^\star}\}, \tag{1}$$

where $a_+ = \max(a, 0)$ for real a and also implies convergence of both the corresponding density function and of its derivative uniformly in y on all finite intervals (Pickands (1986), Theorem 5.2). Thus, if a_m and b_m were known, we might approximate the distribution of $(M_m - b_m)/a_m$ by the right-hand side of (1). In practice, they are unknown, so we fit the generalized extreme value distribution $\mathsf{GEV}(\mu, \sigma, \xi)$ with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma \in \mathbb{R}_+$, and shape parameter $\xi \in \mathbb{R}$,

$$G(y) = \begin{cases} \exp\left\{-\left(1+\xi\dfrac{y-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, & \xi \neq 0, \\ \exp\left\{-\exp\left(-\dfrac{y-\mu}{\sigma}\right)\right\}, & \xi = 0, \end{cases} \tag{2}$$

with this formula valid in $\{y \in \mathbb{R} : \xi(y-\mu)/\sigma > -1\}$. Letting $\xi \to 0$ yields the Gumbel distribution; the limit is continuous, but the log-likelihood and its derivatives should be computed very accurately near $\xi = 0$ to palliate numerical instability of $\log(1+x)$ when $x \approx 0$.

The $\mathsf{GEV}(\mu, \sigma, \xi)$ is max-stable, and this allows extrapolation beyond the observed data into the tail of the distribution: if $Y_1, \ldots, Y_N \sim \mathsf{GEV}(\mu, \sigma, \xi)$ are independent, then $\max\{Y_1, \ldots, Y_N\} \sim \mathsf{GEV}(\mu_N, \sigma_N, \xi)$ with $\mu_N = \mu + \sigma(N^\xi - 1)/\xi$ and $\sigma_N = \sigma N^\xi$ when $\xi \neq 0$, and with $\mu_N = \mu + \sigma \log(N)$ and $\sigma_N = \sigma$ when $\xi = 0$. Thus, if μ, σ and ξ have been estimated from n independent block maxima and the fit of (2) appears adequate, the distribution of a maximum of N further independent observations can also be estimated, even if $N \gg n$, though the uncertainty generally grows alarmingly as N increases.

If equation (1) holds, then the linearly rescaled conditional distribution of an exceedance over a threshold $u < y^*$ also converges (e.g., Embrechts, Klüppelberg and Mikosch (1997), Theorem 3.4.5). Let $r(y) = \{1 - F(y)\}/f(y)$ denote the reciprocal hazard function, then

$$\lim_{u \to y^*} \frac{1 - F\{u + r(u)y\}}{1 - F(u)} = 1 - H(y; 1, \xi), \tag{3}$$

where

$$H(y; \tau, \xi) = \begin{cases} 1 - (1 + \xi y/\tau)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\tau)_+, & \xi = 0, \end{cases} \tag{4}$$

is the generalized Pareto (GP) distribution function with scale $\tau \in \mathbb{R}_+$ and shape $\xi \in \mathbb{R}$, denoted $\mathsf{GP}(\tau, \xi)$. If $Y \sim \mathsf{GP}(\tau, \xi)$, straightforward calculations show that $Y - u \mid Y > u \sim \mathsf{GP}(\tau + \xi u, \xi)$ for any $u \in \mathbb{R}$ such that $\tau + \xi u > 0$, so exceedances above a threshold u also follow a GP distribution. This property is termed threshold-stability, and its consequences parallel those of max-stability.

If the data consist of independent and identically distributed random variables that arrive regularly, for example on a daily basis, and that satisfy the conditions above, then the times of events that exceed the threshold u can be approximated by a homogeneous Poisson process; the exceedances themselves are independent generalized Pareto variables, and this induces a Poisson process of independent event times and sizes $(t_1, y_1), \ldots, (t_n, y_n)$.

The shape parameter ξ determines how the tail probability declines at extreme levels: a negative value gives a distribution with bounded support, whereas zero or positive values give unbounded extremes, with the distribution tail of form $x^{-1/\xi}$ for large x, so larger values of ξ correspond to increasingly heavy tails. Authors who have studied likelihood inference for ξ and other parameters of extreme value distributions include Pires, Cysneiros and Cribari-Neto (2018), who focused on the shape parameter of the generalized Pareto distribution, and Giles, Feng and Godwin (2016) and Roodman (2018), who considered bias-corrected estimates of extreme-value parameters. Although ξ gives important qualitative insights into rare events, it is not a direct measure of risk, and for this reason and because of the previous work we do not consider it to be the primary target of inference here.

2.2. *Risk measures*. Max- and threshold-stability allow the estimation of risks associated with rare events by extrapolating the fits of the GEV or GP distributions. The most common risk measure is the N-year return level, that is, the $(1 - 1/N)$-quantile of F for an annual maximum, often interpreted as "the level exceeded by an annual maximum on average once every N years." The probability p_l that a N-year return level is exceeded l times in N years of independent annual maxima may be computed using a binomial distribution with N trials and success probability $1/N$. For large N, a Poisson approximation yields $p_0 = p_1 = 0.368$, $p_2 = 0.184$, $p_3 = 0.061$, and $p_4 = 0.015$, so the probability of at least one exceedance over N years is, in fact, roughly 0.63. Perhaps more to the point, any return level is a parameter of a distribution. Even if this was known perfectly, risk would be associated with the uncertain nature of future events. In the Bayesian paradigm, one could measure risk using the posterior predictive distribution of the N-year maximum which can be approximated by higher-order

techniques (Davison (1986)). Cox, Isham and Northrop (2002, Section 3b), suggested using direct summaries of the distribution of the N-year maximum also in a frequentist setting; the N-year return level approximates the 0.368 quantile of this distribution.

We mentioned above that the distribution $G_N(y)$ of the maximum of N independent and identically distributed $\mathsf{GEV}(\mu, \sigma, \xi)$ variates is $\mathsf{GEV}(\mu_N, \sigma_N, \xi)$. Denote the expectation and p quantile of the N-year maximum by $\mathfrak{e}_N$ and $\mathfrak{q}_p = G_N^{-1}(p)$ and the associated return level by $z_N = G^{-1}(1 - 1/N)$. These may all be expressed in the form

$$\begin{cases} \mu + \sigma(\kappa_\xi - 1)/\xi, & \xi < 1, \xi \neq 0, \\ \mu + \sigma\kappa_0, & \xi = 0, \end{cases}$$

where κ_ξ equals $N^\xi \Gamma(1 - \xi)$ for $\mathfrak{e}_N$, $\{-N/\log(p)\}^\xi$ for $\mathfrak{q}_p$ and $\{-\log(1 - 1/N)\}^{-\xi}$ for z_N, and κ_0 equals $\log(N) + \gamma_e$ for $\mathfrak{e}_N$, $\log(N) - \log\{-\log(p)\}$ for $\mathfrak{q}_p$ and $-\log\{-\log(1 - 1/N)\}$ for z_N. Inference on any such scalar risk measure ψ is performed by reparametrizing the generalized extreme value distribution in terms of ψ and treating two of the other GEV parameters (μ, σ, ξ) as the vector $\boldsymbol{\lambda}$ of "nuisance" parameters.

Threshold exceedances are related to maxima as follows: suppose we fit a $\mathsf{GP}(\tau, \xi)$ distribution to exceedances above a threshold u, and let ζ_u denote the unknown proportion of points above u. If there are, on average, N_y observations per year, then we take $H^{\zeta_u N N_y}$ as an approximation to the distribution of the N-year maximum above u.

2.3. *Penultimate approximation.* When the GEV or GP distribution is fitted to maxima or threshold exceedances, the best approximating extremal distribution will generally depend on the block size m or threshold u and the shape parameters will differ from the limiting values arising when $m \to \infty$ or $u \to x^*$. Smith (1987) shows that if $\xi^\star = \lim_{m\to\infty} s'(b_m)$ exists, then, for any $x \in \{y : 1 + \xi^\star y > 0\}$, there exists z such that

$$\frac{-\log[F\{v + s(v)x\}]}{-\log\{F(v)\}} = \{1 + s'(z)x\}^{-1/s'(z)}, \quad v < z < v + s(v)x.$$

For each $m \geq 1$, setting $v = b_m$, and $a_m = s(b_m)$ yields

$$F^m(a_m x + b_m) = \exp\big[-\{1 + s'(z)x\}^{-1/s'(z)}\big] + \mathrm{O}(m^{-1})$$

which can be regarded as a finite-m, or penultimate, version of the approximation stemming from equation (1). Smith shows that the Hellinger distance between $F^m(a_m x + b_m)$ and the penultimate approximation $\mathsf{GEV}\{0, 1, s'(b_m)\}$ approaches zero as $m \to \infty$ and that it is smaller than that between $F^m(a_m x + b_m)$ and $\mathsf{GEV}(0, 1, \xi^\star)$. Similar statements hold for the generalized Pareto distribution: unless $r'(x)$ is constant, there exists y such that a finite-u version of equation (3),

$$\frac{1 - F\{u + r(u)x\}}{1 - F(u)} = \{1 + r'(y)x\}_+^{-1/r'(y)}, \quad u < y < u + r(u)x,$$

holds. One can replace the limiting shape $\lim_{v\to x^*} r'(v) = \xi^\star$ by $r'(u)$, thereby reducing the Hellinger distance between the true conditional distribution of exceedances and the generalized Pareto approximation. Penultimate appproximations for specific parametric models are straightforward to obtain, as one only needs to compute the scale a_m, location b_m, and shape $s'(b_m)$ parameters for the GEV approximation or the scale $r(u)$ and shape $r'(u)$ parameters for the GP approximation.

When the limiting parametric models are fitted to finite samples, maximum likelihood estimates of the shape parameter will tend to be closer to their penultimate counterparts than to $\xi^\star$; moreover, the estimator of ξ will converge to a target that changes as the threshold or

the block size increases, depending on the curvature of s' or r'. Extrapolations far beyond the data will, inevitably, be biased due to the incorrect assumption that the max- or threshold-stability property that holds for infinite m or the limiting threshold also applies at finite levels. It is customary to investigate the properties of estimators of risk or of ξ by comparing them with their asymptotic limits, but their penultimate counterparts may be better finite-sample targets of inference.

2.4. *Finite-sample bias.* Although many approaches to estimation of the generalized Pareto and generalized extreme-value distribution have been proposed, we shall consider likelihood-based estimation, which can easily be extended to complex settings and sampling schemes. Consistency and asymptotic normality of maximum likelihood estimators for the extremal distributions have been established (Bücher and Segers (2017); Dombry and Ferreira (2019), and references therein), but such studies consider infinite sample size, while small-sample biases can arise, even when the assumed model is correct.

The finite-sample properties of maximum likelihood estimators for extreme value distributions can be poor (e.g., Hosking and Wallis (1987), Table 5), due to their small-sample bias (Giles, Feng and Godwin (2016); Roodman (2018)). Figure 2 illustrates this for the shape parameter ξ. Apart from becoming more concentrated as the sample size n increases, the distribution of $\widehat{\xi} - \xi$ when fitting the GEV distribution to maxima depends little on n, and, in particular, its median barely changes. By contrast, the distribution of $\widehat{\xi} - \xi$, based on the GP distribution, shows strong negative skewness, with its median and lower quartile systematically increasing as n increases, while the upper quartile barely changes. It turns out that the scale estimator $\hat{\sigma}$ for the GP distribution is upwardly biased, and this partially compensates for the downward bias of $\widehat{\xi}$, but extrapolation far into the tail based on a GP fit tends to underestimate the sizes of extreme events anyway. Bias-correction can mitigate this, but analytical first-order corrections of the type pioneered by Cox and Snell (1968) are applicable only when $\xi < -1/3$ and are unbounded near this. Analytical bias correction is quite different from bootstrap bias correction (Belzile (2019)), and, furthermore, bias-correction formulae for risk measures are currently unavailable. We thus consider implicit bias corrections below.

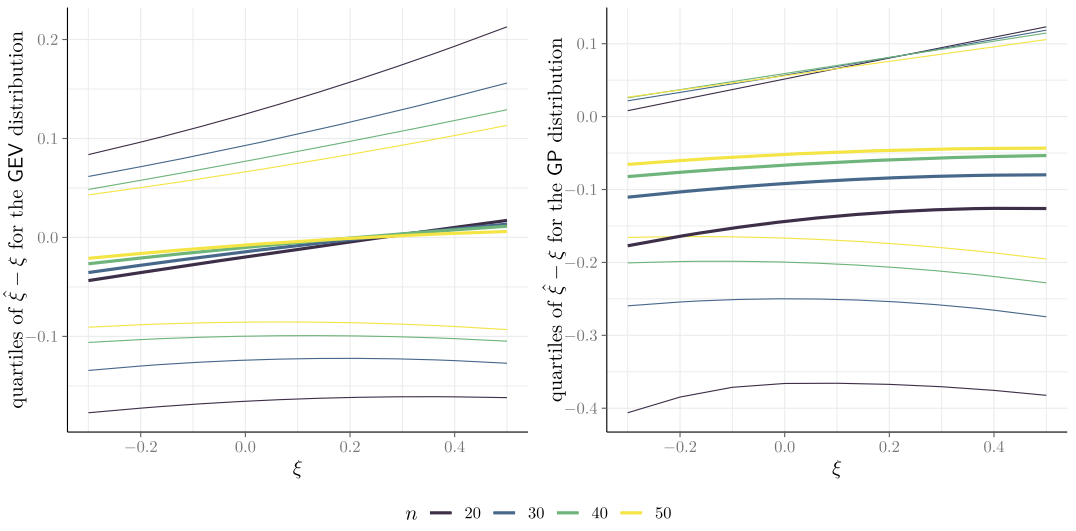

FIG. 2. *Smoothed quartiles of the distribution of differences between maximum likelihood estimates and true shape parameter,* $\widehat{\xi} - \xi$, *based on* 13,000 *simulations from* $\mathsf{GEV}(0, 1, \xi)$ *and* $\mathsf{GP}(1, \xi)$ *distributions for samples of sizes* $n = 20, 30, 40, 50$.

3. Inference.

3.1. *Likelihood.* Consider a parametric model for observations $y_1, \ldots, y_n$ with log-likelihood function $\ell(\boldsymbol{\theta})$ whose p-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ can be decomposed into a q-dimensional parameter of interest $\boldsymbol{\psi}$ and a $(p-q)$-dimensional nuisance vector $\boldsymbol{\lambda}$. The score vector $U(\boldsymbol{\theta})$, the observed information matrix $j(\boldsymbol{\theta})$, and its inverse are partitioned accordingly, and the maximum likelihood estimate is $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}})$.

The profile log-likelihood for the parameter of interest is

$$\ell_{\mathsf{p}}(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda}} \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \ell(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}}) = \ell(\boldsymbol{\psi}, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}).$$

The asymptotic properties of this and related statistics stem from those of the full likelihood and are standard under mild conditions (cf. Severini (2000), page 128): for example, if ψ is scalar with true value ψ_0, the likelihood root

$$R(\psi_0) = \operatorname{sign}(\widehat{\psi} - \psi_0)\big[2\{\ell_{\mathsf{p}}(\widehat{\psi}) - \ell_{\mathsf{p}}(\psi_0)\}\big]^{1/2} \tag{5}$$

has an asymptotic standard normal distribution to order $\mathrm{O}(n^{-1/2})$. Confidence limits ψ_α for ψ are obtained by solving the equations $R(\psi_\alpha) = \Phi^{-1}(\alpha)$ for $\alpha \in (0,1)$, with Φ denoting the standard normal cumulative distribution function. Such intervals are invariant to interest-preserving reparametrisations. The same two-sided equitailed confidence intervals are obtained via χ^2 approximation to the distribution of the likelihood ratio statistic based on $\ell_{\mathsf{p}}(\psi)$, but these and other first-order methods can perform badly in small samples if the dimension of $\boldsymbol{\lambda}$ is large, and higher-order methods may then provide more accurate tests and confidence intervals.

The standard asymptotic approximations described above apply under regularity conditions whereby the score statistic satisfies the first two Bartlett equalities. The rth moment of the score for the GEV and GP distributions exists only for $\xi > -1/r$ (Smith (1985)), so the above discussion applies to them only for $\xi > -1/2$. Estimates of ξ are typically above -0.3 in applications and very often are close to zero, so the failure of regularity conditions is rarely of practical importance.

In the case of scalar ψ, one improvement is via normal approximation to a modified likelihood root (Barndorff-Nielsen and Cox (1994), Section 6.6.1)

$$R^\star(\psi) = R(\psi) + \frac{1}{R(\psi)} \log\left\{\frac{Q(\psi)}{R(\psi)}\right\}, \tag{6}$$

where $Q(\psi)$ is discussed below. If the response distribution is continuous, then $R^\star(\psi_0)$ is asymptotically standard normal to order $\mathrm{O}(n^{-3/2})$; this is known as a third-order approximation. In many ways more important than the reduction of the error rate from $n^{-1/2}$ to $n^{-3/2}$ is the fact that the error when using $R^\star(\psi)$ is relative, leading to improved inferences, even when $\widehat{\psi}$ is distant from ψ_0. Confidence limits are obtained by solving the equations $R^\star(\psi_\alpha) = \Phi^{-1}(\alpha)$, and these too are invariant to interest-preserving reparametrization to the given order.

Estimators of ψ can be obtained by solving the equations $R(\psi) = 0$ and $R^\star(\psi) = 0$. The first yields the maximum likelihood estimator $\widehat{\psi}$, while the second yields $\widehat{\psi}^\star$, an implicitly debiased version of $\widehat{\psi}$; both transform appropriately. The maximum likelihood estimator could also be debiased directly by subtracting an estimated bias or, indirectly, by modifying the corresponding score equation (Firth (1993); Kenne Pagui, Salvan and Sartori (2017); Belzile (2019)).

The use of equation (6) hinges on the ready computation of $Q(\psi)$. This involves sample space derivatives of the log-likelihood, which are awkward, in general, and a variety of approaches to their computation have been proposed. In the next section we sketch a simple general approach developed by D. A. S. Fraser, N. Reid, and colleagues.

3.2. *Tangent exponential model*. The tangent exponential model (TEM) provides a general formula for the quantity $Q(\psi)$ that appears in equation (6) (Fraser, Reid and Wu (1999)). The idea is to approximate the probability density function of the data by that of an exponential family for which highly accurate inference is possible. Following Brazzale, Davison and Reid (2007, Chapter 8), we outline its construction. The presence of sample space derivatives makes it necessary to distinguish a generic response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ from the responses actually observed, $\boldsymbol{y}^{\mathrm{o}} = (y_1^{\mathrm{o}}, \ldots, y_n^{\mathrm{o}})^\top$.

The tangent exponential model depends on an $n \times p$ matrix $\mathbf{V}$, whose ith row equals the derivative of y_i with respect to $\boldsymbol{\theta}^\top$, evaluated at $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{y}^{\mathrm{o}}$; the p columns of $\mathbf{V}$ correspond to vectors in $\mathbb{R}^n$ that are informative about the variation of $\boldsymbol{y}$ with $\boldsymbol{\theta}$. The TEM implicitly conditions on an $(n-p)$-dimensional approximate ancillary statistic whose value lies in the space orthogonal to the columns of $\mathbf{V}$ and constructs a local exponential family approximation at $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{y}^{\mathrm{o}}$ with canonical parameter

$$\boldsymbol{\varphi}(\boldsymbol{\theta}) = \mathbf{V}^\top \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{y}}\bigg|_{\boldsymbol{y}=\boldsymbol{y}^{\mathrm{o}}}.$$

The components of $\boldsymbol{\varphi}(\boldsymbol{\theta})$ can be interpreted as the directional derivatives of $\ell(\boldsymbol{\theta}; \boldsymbol{y}^{\mathrm{o}} + \mathbf{V}\boldsymbol{t})$ with respect to the columns of $\mathbf{V}$, obtained by differentiating with respect to the components of the $p \times 1$ vector $\boldsymbol{t}$ and setting $\boldsymbol{t} = \mathbf{0}$.

In models for continuous scalar responses, the $\mathbf{V}_i$ can be obtained by using the probability integral transform to write $y_i = F^{-1}(u_i; \boldsymbol{\theta})$ in terms of a uniform variable u_i, yielding

$$\mathbf{V}_i = \frac{\partial y_i}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{y}=\boldsymbol{y}^{\mathrm{o}}, \boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{\partial F(y_i^{\mathrm{o}}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \frac{1}{f(y_i^{\mathrm{o}}; \boldsymbol{\theta})}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \quad i = 1, \ldots, n; \tag{7}$$

equivalently, we may take the total derivative of the pivotal quantity $F(y_i; \boldsymbol{\theta})$. Discrete responses cannot be differentiated and are replaced by their means, leading to an error of order n^{-1} for inferences based on equation (6) (Davison, Fraser and Reid (2006)).

The approximate pivot in equation (6) stemming from the TEM is

$$Q(\psi) = \frac{\left|\boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}_\psi) \quad \partial \boldsymbol{\varphi}/\partial \boldsymbol{\lambda}^\top(\widehat{\boldsymbol{\theta}}_\psi)\right|}{|\partial \boldsymbol{\varphi}/\partial \boldsymbol{\theta}^\top(\widehat{\boldsymbol{\theta}})|} \frac{|j(\widehat{\boldsymbol{\theta}})|^{1/2}}{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_\psi)|^{1/2}}, \tag{8}$$

where the first matrix in the numerator is formed by binding the $p \times 1$ vector $\boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}_\psi)$ to the $p \times (p-1)$ matrix $\partial \boldsymbol{\varphi}/\partial \boldsymbol{\lambda}^\top$. A modified profile likelihood $\ell_{\mathrm{fr}}(\psi) = -\{R^\star(\psi)\}^2/2$ may be constructed by using equation (8) and treating $R^\star(\psi)$ as standard normal.

3.3. *Modified profile likelihoods*. In the previous section the likelihood root was derived from the profile log-likelihood function via the likelihood ratio statistic, then modified, and used to construct the modified profile likelihood $\ell_{\mathrm{fr}}(\psi)$. An alternative is direct modification of the profile log-likelihood, two approaches to which are listed by Severini (2000, Sections 9.5.3–9.5.4). The first approach uses elements of the tangent exponential model and is of the form

$$\ell_{\mathrm{m}}^{\mathrm{tem}}(\boldsymbol{\psi}) = \ell_{\mathrm{p}}(\boldsymbol{\psi}) + \frac{1}{2}\log\{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})|\} - \log\{|\ell_{\boldsymbol{\lambda};\boldsymbol{y}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\mathbf{V}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})|\}, \tag{9}$$

where $\ell_{\boldsymbol{\lambda};\boldsymbol{y}} = \partial^2 \ell/\partial \boldsymbol{\lambda} \partial \boldsymbol{y}^\top$ is the derivative of the log-likelihood with respect to the nuisance parameter and observations and $\mathbf{V}_{\boldsymbol{\lambda}}$ denotes the columns of $\mathbf{V}$ in equation (7) that correspond to derivatives with respect to $\boldsymbol{\lambda}$. The second approach is due to Severini and is similar in spirit to ideas in Skovgaard (1996). This approximation has a lower order of accuracy than the TEM and relies on empirical covariances with

$$\ell_{\mathrm{m}}^{\mathrm{cov}}(\boldsymbol{\psi}) = \ell_{\mathrm{p}}(\boldsymbol{\psi}) + \frac{1}{2}\log\{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})|\} - \log\{|\widehat{j}_{\boldsymbol{\lambda};\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}}; \widehat{\boldsymbol{\theta}})|\}, \tag{10}$$

where

$$\widehat{j}_{\boldsymbol{\lambda};\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}};\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \ell_{\boldsymbol{\lambda}}^{(i)}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})\ell_{\boldsymbol{\lambda}}^{(i)}(\widehat{\boldsymbol{\theta}})^{\top};$$

here, $\ell_{\boldsymbol{\lambda}}^{(i)}$ denotes the component of the score statistic due to the ith observation. It is straightforward to check that these modifications are invariant to interest-preserving reparametrisation. The components of the score vector and information matrix are readily obtained, if necessary using computer algebra.

4. Data analyses.

4.1. *Vargas tragedy.* Figure 1 only considered threshold exceedances of events before the catastrophe, either treating the number of threshold exceedances n_u as fixed, incorrectly, or as random. To deepen our analysis, we consider how risk measures are affected by stopping data collection after the disaster and add the yearly daily rainfall maxima for 1951–1960 to the daily values for 1961–1999, considered in Section 1.2. Failing to account for the fact that the data collection appears to end with the largest event ever observed leads to upwardly-biased risk estimates (Barlow, Sherlock and Tawn (2020)). There are a number of ways in which an implicit stopping rule might be formulated, but for a rough assessment of its impact we consider that sampling would have ended at the first time a daily value exceeded $s = 282$ mm, the largest two-day sum previously reported.

We consider a Poisson process $\mathcal{P}$ with measure ν on $\mathcal{X} = (0, \infty) \times (u, \infty)$, where the first axis represents the times t of extreme events and the second axis represents their sizes y, which are presumed to exceed some high threshold u. If the events $x = (t, y)$ arrive at a constant rate in time, then standard extremal modeling leads to setting

$$\nu\{[t_1, t_2] \times (y, \infty)\} = (t_2 - t_1)\Lambda(y), \tag{11}$$

where $\Lambda(y) = \{1 + \xi(y - \mu)/\tau\}_+^{-1/\xi}$ and the intensity of the process is $\dot{\nu}(t, y) = -\dot{\Lambda}(y) = -\partial\Lambda(y)/\partial y$, say. Our stopping rule presupposes that for some given $s > u$, $\mathcal{X}$ is partitioned into a stopping set $\mathcal{S} = (0, \infty) \times (s, \infty)$ and its complement $\mathcal{S}^{\mathsf{c}}$, and the data are analysed just after the random time T at which $\mathcal{P}$ first falls into $\mathcal{S}$. The probability of no event in $\mathcal{S}$ before time t is $\exp\{-t\Lambda(s)\}$, so T is an exponential random variable with mean $\Lambda(s)^{-1}$. Let N_t denote the number of events in $\mathcal{S}^{\mathsf{c}}$ before time t. Since $\mathcal{S}$ and $\mathcal{S}^{\mathsf{c}}$ are disjoint, events in them are independent, so T is independent of events in $\mathcal{S}^{\mathsf{c}}$. Hence the probability element corresponding to successive events $(t_1, y_1), \ldots, (t_n, y_n)$ in $\mathcal{S}^{\mathsf{c}}$ followed by (t, y_t) in $\mathcal{S}$ is

$$\dot{\nu}(t, y_t) \prod_{j=1}^{n} \dot{\nu}(t_j, y_j) \times \exp[-\nu\{[0, t] \times [u, \infty)\}], \quad u < y_1, \ldots, y_n < s < y_t. \tag{12}$$

The stopping rule affects the repeated sampling properties of estimators by constraining the data to satisfy $N_t = n$. We can rewrite (12) as

$$\exp\big[-t\{\Lambda(u) - \Lambda(s)\}\big] \prod_{j=1}^{n} \{-\dot{\Lambda}(y_j)\} \times \exp\{-t\Lambda(s)\}\{-\dot{\Lambda}(y_t)\}, \tag{13}$$

where the first term corresponds to the n events in $\mathcal{S}^{\mathsf{c}}$ before time t and the second term to the event in $\mathcal{S}$ that terminates the sampling. In the application to the Vargas data, we take $t = 39$ years, so an annual maximum has distribution $\exp\{-\Lambda(y)\}$. The probability that sampling stops at time t and there are then n events in $\mathcal{S}^{\mathsf{c}}$ is

$$\mathsf{P}(N_t = n, T = t) = \frac{[t\{\Lambda(u) - \Lambda(s)\}]^n}{n!} \exp\big[-t\{\Lambda(u) - \Lambda(s)\}\big] \times \exp\{-t\Lambda(s)\} \times \Lambda(s).$$

Expression (13) corresponds to what Barlow, Sherlock and Tawn (2020) call $\mathscr{L}_{\text{std}}$, and which they find gives biased inferences and the equivalent of their "full conditional" likelihood, which uses the joint density of the event sizes $y_1, \ldots, y_n, y_t$, conditional on $T = t$ and $N_t = n$, that is,

$$\mathscr{L}_{\text{fc}} = \left\{\frac{-\dot{\Lambda}(y_t)}{\Lambda(s)}\right\} \prod_{j=1}^{n} \left[\frac{-\dot{\Lambda}(y_j)}{t\{\Lambda(u) - \Lambda(s)\}}\right], \quad u < y_1, \ldots, y_n < s < y_t, \tag{14}$$

should have lower sampling bias. If $y > u$, then $-\dot{\Lambda}(y)/\Lambda(u)$ equals a generalized Pareto density with parameters ξ and $\tau_u = \sigma + \xi(u - \mu)$ evaluated at $y - u$, so (14) reduces to a product of truncated generalized Pareto densities. We do not consider the "partial conditional" likelihood by Barlow, Sherlock and Tawn (2020) of $\mathscr{L}_{\text{pc}}$, which corresponds to replacing the terms $\Lambda(u) - \Lambda(s)$ in (14) by $\Lambda(u)$, that is, ignoring the right truncation of $y_1, \ldots, y_n$.

Our functional of interest, the median of the semicentennial daily maximum, is obtained by taking $N = 50$ years, $p = 0.5$, and

$$\mathfrak{q}_p = \mu - \frac{\sigma}{\xi}\left\{1 - \left(\frac{-N}{\log p}\right)^{\xi}\right\}.$$

The log-likelihood has three components: the yearly maxima for 1951–1960, the right-truncated exceedances of $u = 27$ mm for 1961–1999 and, finally, a left-truncated largest record that exceeds the stopping rule threshold $s = 282$ mm. As these are independent, they make additive contributions to $\boldsymbol{\varphi}(\boldsymbol{\theta})$. The component for the maxima is readily obtained using their assumed generalized extreme-value distribution and so are the components for the exceedances in the full conditional likelihoods which correspond to independent truncated generalized Pareto variables; see Appendix C. The form of $\boldsymbol{\varphi}(\boldsymbol{\theta})$ for the tangent exponential model approximation for a Poisson process, which is needed for the higher-order version of (13), seems not to have been derived previously and may be found in Appendix B.

The ordinary and TEM-based profile log likelihoods for $\mathfrak{q}_p$ with full conditioning and the standard log-likelihood, shown in the left panel of Figure 3, are strikingly different: the conditional log-likelihood is much more concentrated and the higher-order TEM version is just

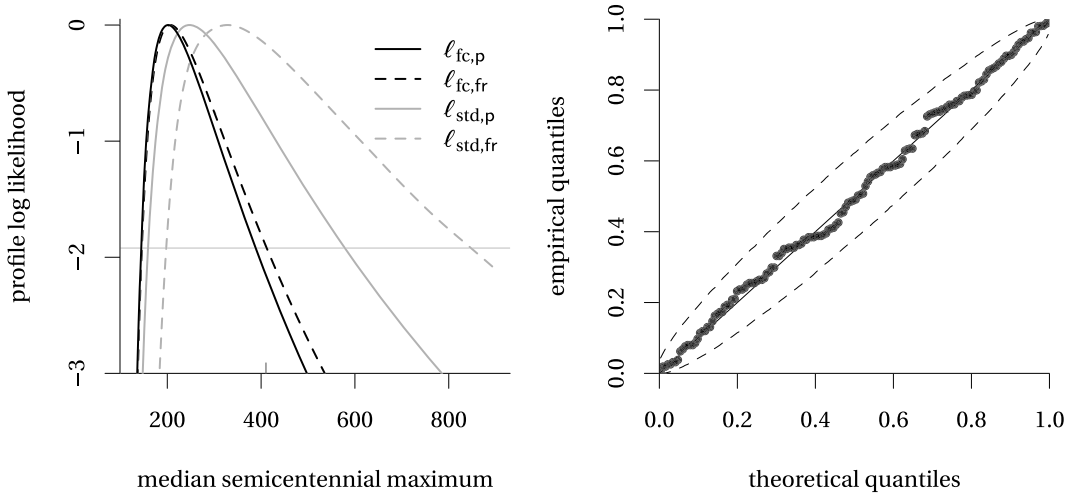

FIG. 3. *Left panel*: *Profile log-likelihood based on* $\mathscr{L}_{std}$ (*gray*) *and* $\mathscr{L}_{fc}$ (*black*) *for the Maiquetía data using threshold exceedances up to and including data for December* 15*th*, 1999. *The curves show the shifted regular profile likelihood* (*full*) *and the* TEM *approximation* $\ell_{\text{fr}} = -R^{\star}(\psi)^2/2$ (*dashed*). *The dashed grey horizontal line at* -1.92 *indicates cutoff values for* 95% *confidence intervals based on the asymptotic* χ^2_1 *distribution. The mark at* 410.4 *mm indicates the record of December* 15*th*, 1999. *Right panel*: *Probability-probability plot for the full conditional likelihood fit with approximate simultaneous* 95% *confidence intervals*.

slightly less precise, whereas the TEM version of the standard log-likelihood (13) gives much larger point estimates and upper confidence interval limit than does the standard profile log-likelihood. Thus, not allowing for the implicit stopping rule can have a dramatic effect not only on standard but also on higher-order inferences. If we consider the one-sided likelihood ratio test for $\mathscr{H}_0 : \mathfrak{q}_p > 410.4$ mm, the respective p-values obtained from $R/R^\star$ are 0.095/0.28 for the standard likelihood and 0.019/0.025 for the full conditional likelihood, suggesting that the magnitude of the 1999 event was indeed significantly larger than the median 50-year maximum.

4.2. *Venice sea level.* The Italian city of Venice is threatened by sea-level rise and subsidence and is increasingly at risk from flooding in so-called *acqua alta* events. To quantify this risk, we consider data analyzed by Smith (1986) and Pirazzoli (1982) containing large annual sea level measurements from 1887 until 1981, complemented with series for 1982–2019 extracted from the City of Venice website (accessed June 2020 and available under the CC BY-NC-SA 3.0 license). Only the yearly maximum is available for 1922 and only the six largest observations for 1936. Figure 4 shows the two largest annual order statistics; while there is a clear trend, we detected no change when the measurement gauge was relocated in 1983. In addition to the simple straight-line model suggested by the plot, we fitted a smooth additive nonparametric quantile regression (Fasiolo et al. (2021)) with 50 knots and a smooth term for years and different intercepts for the two largest order statistics: the resulting fits, shown in Figure 4, suggest that a straight line is adequate.

If the extremal types theorem holds, then the log-likelihood corresponding to the joint limiting distribution of the r largest observations of a sample, $Y_1 \geq \cdots \geq Y_r$, is

$$\ell(\mu, \sigma, \xi; \boldsymbol{y}) = -r\log(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{r} \log\left(1 + \xi \frac{y_j - \mu}{\sigma}\right)_+ - \left(1 + \xi \frac{y_r - \mu}{\sigma}\right)_+^{-1/\xi}, \quad \mu, \xi \in \mathbb{R}, \sigma > 0. \tag{15}$$

The 10 largest sea levels are available for almost every year, but one might ask whether they should be used. The model presupposes that they arise from independent underlying

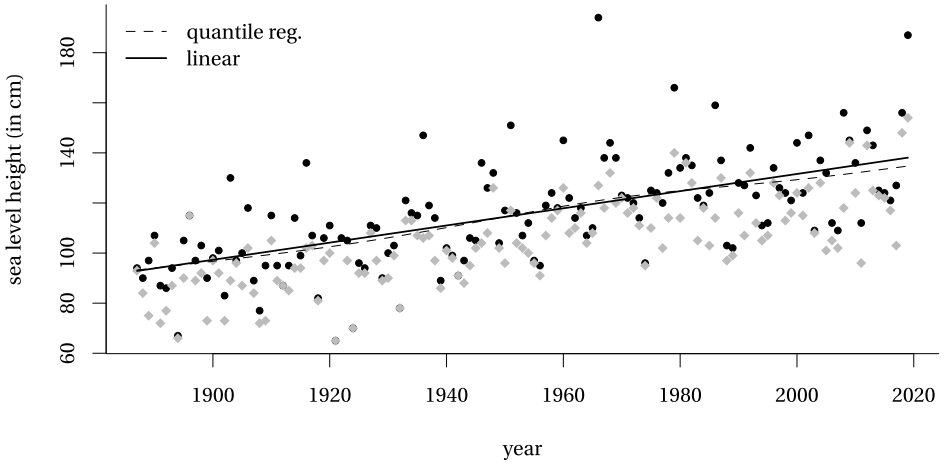

FIG. 4. *First* (*black*) *and second* (*grey*) *largest yearly observations for the Venice sea level data* (*in cm*), *a smooth additive quantile regression model for the median* (*with a smooth term for time and a different intercept for each order statistic*). *The predicted values for the largest order statistic and those for the corresponding linear regression are superimposed.*

variables, but, in practice, many are due to combinations of high tides and bad weather during the winter months. The data source for recent years allows apparently independent events to be identified, but this is harder for the earlier data.

One purely statistical basis for choosing r is by balancing the information added as r increases against the potential for bias when r is too large. Calculations in Appendix A establish that the 3×3 Fisher information matrix, based on (15), is of the form $I_r(\mu, \sigma, \xi) + (r-1)I(\mu, \sigma, \xi)$, where $I_r(\mu, \sigma, \xi)$ stems from Y_r, and $(r-1)I(\mu, \sigma, \xi)$ is the contribution for the other observations. These matrices can be used to compute the information gained when basing inference on $Y_1, \ldots, Y_r$ rather than only on the sample maximum, Y_1. To do so, we calculate the ratios of the diagonal elements of $I_1^{-1}(\mu, \sigma, \xi)$ to those of $\{I_r(\mu, \sigma, \xi) + (r-1)I(\mu, \sigma, \xi)\}^{-1}$; an overall variance reduction for a given r is

$$\left\{ \frac{|I_1(\mu, \sigma, \xi)|}{|I_r(\mu, \sigma, \xi) + (r-1)I(\mu, \sigma, \xi)|} \right\}^{1/3}.$$

Figure 5 shows the variance reduction factors for μ, σ, ξ, and the overall efficiency. There seems to be little gain from taking $r > 5$ for estimation of μ and σ, while for ξ the decline is closer to that of independent generalized extreme value data. This is because the parameters μ and σ cannot be estimated based only on $I(\mu, \sigma, \xi)$, which has rank two, whereas both $I(\mu, \sigma, \xi)$ and $I_r(\mu, \sigma, \xi)$ contain information on ξ. Hence, as r increases, the information gain for the location and scale parameters becomes more limited.

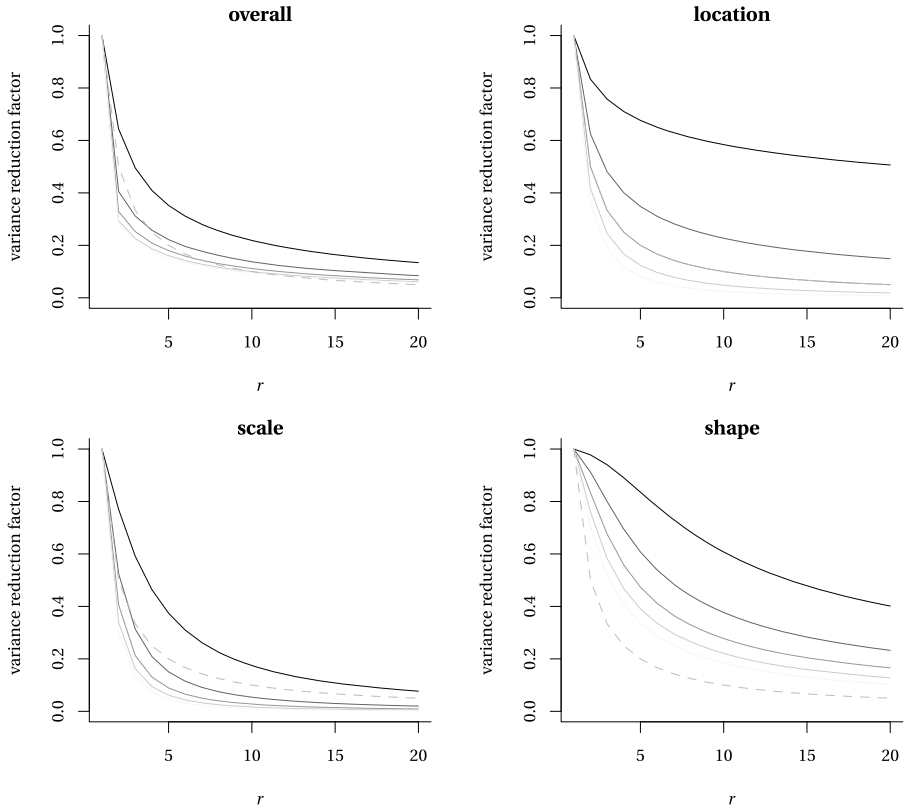

FIG. 5. *Variance reduction factors for inference based on the r-largest order statistics model for the location, shape and scale parameters, and overall efficiency (clockwise from top right), as functions of r. The dashed grey line shows the efficiency gain for independent observations. The value of the shape parameter ranges from* $\xi = -0.4$ *(full black) to* $\xi = 0.4$ *(full pale grey) in increments of* 0.2.

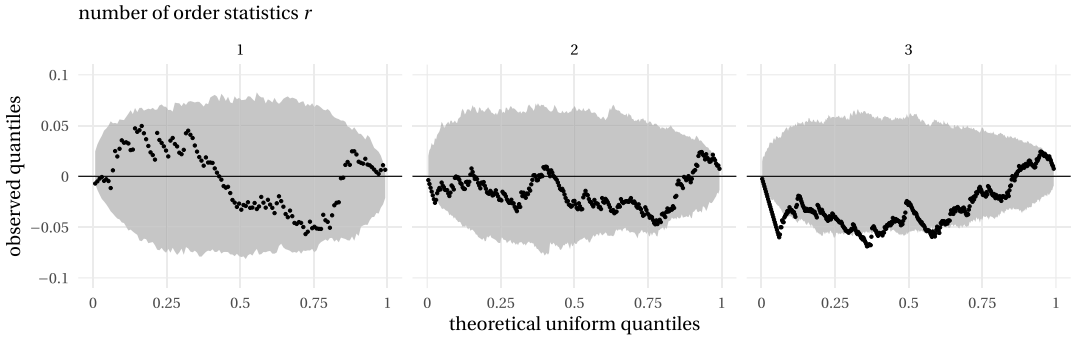

FIG. 6. *Tukey's detrended probability-probability plot for the spacings of the r-largest order statistics of the Venice data, with r = 1, 2, 3, with approximate simultaneous 95% confidence intervals obtain by using a parametric bootstrap and the envelope method (Davison and Hinkley (1997), Section 4.2.4).*

The fit can be checked by noting that if the model is correct, then

$$0 < \Lambda_{\boldsymbol{\theta}}(y_1) < \Lambda_{\boldsymbol{\theta}}(y_2) < \cdots < \Lambda_{\boldsymbol{\theta}}(y_r), \quad \Lambda_{\boldsymbol{\theta}}(y) = \{1 + \xi(y-\mu)/\sigma\}_+^{-1/\xi} \tag{16}$$

are a realisation of the first r points of a unit rate Poisson process on the positive half-line. This implies that the spacings $\Lambda_{\boldsymbol{\theta}}(y_{(1)})$, $\Lambda_{\boldsymbol{\theta}}(y_2) - \Lambda_{\boldsymbol{\theta}}(y_1)$, ... have standard exponential distributions, and systematic departures from this will indicate model failure. The r largest observations from the asymptotic model can be generated by simulating a unit rate Poisson process $0 < U_1 < U_2 < \cdots$, where $U_j = E_1 + \cdots + E_j$ and $E_j \sim \mathsf{Exp}(1)$, and setting $Y_j = \mu + \sigma(U_j^{-1/\xi} - 1)/\xi$. The estimated inverse transformation $\Lambda_{\widehat{\boldsymbol{\theta}}}$ can be used to obtain empirical spacings. These should be approximately independent and can be used to construct probability-probability plots, such as Figure 6. The spacings for $r = 3$ are suggestive of model misspecification for the Venice data, so it seems that just two extrema each year should be used. The initial linear decreases visible in the central and right-hand panels of Figure 6 are probably due to ties or near-ties for the lower records, due to rounding of the data.

Below, we use the $r = 2$ largest observations for each year and treat data for different years as independent. Our chosen risk measure is the probability that in year t the annual maximum sea level exceeds the level $z = 194$ cm reached in the catastrophic flooding of 1966, based on a nonstationary extremal model with location parameter $\mu_0 + \mu_1 \mathtt{year}$, σ, and ξ.

In order to compute the terms necessary for the TEM approximation, suppose that we have data $(y_1, \ldots, y_r)$ and pivots

$$u_1(y_1; \boldsymbol{\theta}), u_2(y_1, y_2; \boldsymbol{\theta}), \ldots, u_r(y_1, \ldots, y_r; \boldsymbol{\theta}).$$

Total differentiation of $u_1(y_1; \boldsymbol{\theta})$ yields

$$0 = \frac{\partial u_1(y_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\partial y_1}{\partial \boldsymbol{\theta}} \frac{\partial u_1(y_1; \boldsymbol{\theta})}{\partial y_1},$$

and, therefore,

$$\frac{\partial y_1}{\partial \boldsymbol{\theta}} = -\left\{\frac{\partial u_1(y_1; \boldsymbol{\theta})}{\partial y_1}\right\}^{-1} \frac{\partial u_1(y_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Total differentiation of $u_j(y_1, \ldots, y_j; \boldsymbol{\theta})$ likewise yields

$$\frac{\partial y_j}{\partial \boldsymbol{\theta}} = -\left\{\frac{\partial u_j(y_1, \ldots, y_j; \boldsymbol{\theta})}{\partial y_j}\right\}^{-1} \left\{\frac{\partial u_j(y_1, \ldots, y_j; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_{i=1}^{j-1} \frac{\partial y_i}{\partial \boldsymbol{\theta}} \frac{\partial u_i(y_1, \ldots, y_i; \boldsymbol{\theta})}{\partial y_i}\right\}$$

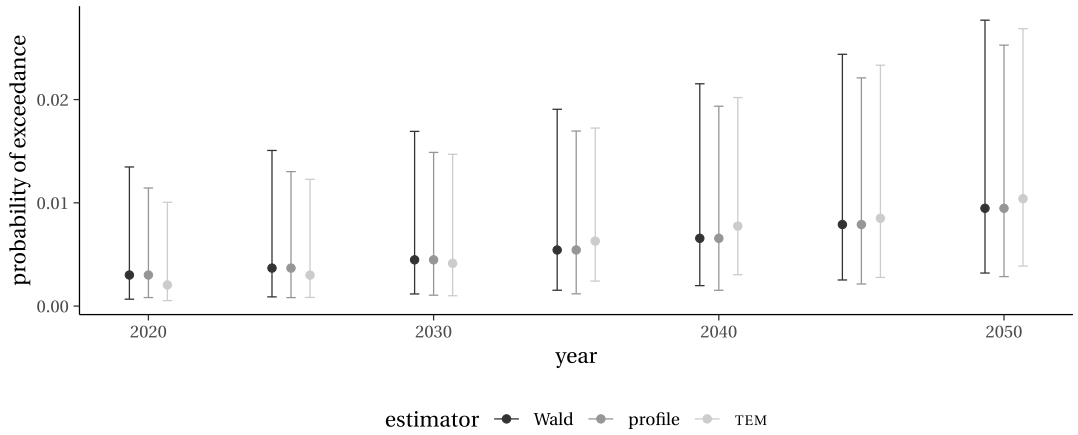

FIG. 7. *Probability of exceedance of* 194 *cm with* 95% *pointwise confidence based on Wald* (*logit-scale*), *likelihood root* R, *and modified likelihood root* $R^\star$ *statistics*.

with all these expressions evaluated at $y_1^{\mathrm{o}}, \ldots, y_j^{\mathrm{o}}$ and $\widehat{\boldsymbol{\theta}}$. In the present case the differences in (16) are pivots, $u_j(y_1, \ldots, y_j; \boldsymbol{\theta}) = \Lambda_{\boldsymbol{\theta}}(y_j) - \Lambda_{\boldsymbol{\theta}}(y_{j-1})$, and the resulting expressions for $\partial y_j / \partial \boldsymbol{\theta}$ involve at most two of the y_i.

Figure 7 shows that the profile- and TEM-based point estimates and 95% confidence intervals for the probability of a flood exceeding the 1966 level for various years are quite similar, though the higher-order estimates vary slightly more over time; similar results are obtained for the other methods of Section 3.3 (not shown). The Wald-based confidence intervals, computed on the logit scale and back-transformed, are somewhat wider. Despite the increase in sea level, it appears that, even without interventions, an event as rare as that in 1966 will remain unlikely for at least the next two decades. The recent inauguration of the Mose system of flood barriers, which can be raised in order to prevent Venice from flooding when there are adverse tides in the Adriatic sea, should reduce this probability yet further, at least in the medium term.

4.3. *Old age in Italy.* The existence or not of a finite upper limit for human lifetimes has recently sparked interest in the extreme value community (Belzile et al. (2021) and (2022); Hanayama and Sibuya (2016); Rootzén and Zholud (2017); Einmahl, Einmahl and de Haan (2019)). The Italian centenarian data set, kindly provided by Holger Rootzén, contains the birth dates and ages of 3836 individuals from a study of semisupercentenarians conducted by the Istituto Nazionale di Statistica (Istat); see Barbi et al. (2018). Individuals are included if they were aged 105 years or more at some point between January 1, 2009 (c_1) and January 1, 2016 (c_2); the survival time is censored for individuals alive at c_2. The cohort comprises persons born between 1896 and 1910 with excess lifetimes measured in days above 105 years, that is, above $u = 38{,}351$ days. It is natural to fit the generalized Pareto model to these excess lifetimes, but it is important to account for the potential left-truncation and right-censoring. Failure to account for the censoring would lead to negative bias for the shape parameter ξ, for example, since individuals born after 1910 could not attain 116 years. A negative shape parameter corresponds to a finite upper limit $\iota = -\tau/\xi$, whereas $\xi \geq 0$ means there is no upper limit.

We consider excess lifetimes of those individuals whose age exceeded u between calendar times c_1 and c_2. Let S and f denote the survival and the density functions of lifetimes, let x_i denote the calendar date at which individual i reached u years, let t_i denote the excess lifetime above u at calendar time c_2, and let a_i be an indicator variable taking value 1 if

TABLE 1

Maximum likelihood estimates of the generalized Pareto for the Italian semisupercentenarian data. From left to right, threshold u (in years), number of threshold exceedances n_u, estimates (standard errors) of the scale and shape parameters σ and ξ, log-likelihood at MLE $\ell(\widehat{\theta})$

u	n_u	$\widehat{\sigma}$	$\widehat{\xi}$	$\ell(\widehat{\boldsymbol{\theta}})$
105	3836	1.67 (0.04)	−0.04 (0.02)	−4253.7
106	1874	1.70 (0.06)	−0.07 (0.03)	−2064.3
107	946	1.47 (0.08)	−0.02 (0.04)	−999.3
108	415	1.47 (0.11)	−0.01 (0.06)	−440.6
109	198	1.33 (0.15)	0.03 (0.09)	−202.9
110	88	1.22 (0.23)	0.12 (0.17)	−85.4
111	34	1.50 (0.47)	0.06 (0.30)	−34.9

individual i was alive at calendar time c_2 and zero otherwise. The resulting likelihood is

$$L(\boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{s}) = \prod_{i=1}^{n} \left[\frac{f(t_i)}{S\{(c_1 - x_i)_+\}} \right]^{1-a_i} \left[\frac{S(t_i)}{S\{(c_1 - x_i)_+\}} \right]^{a_i},$$

with the first and second terms in the product corresponding to those individuals seen to die and to those whose lifetimes are censored at c_2. We fit a generalized Pareto distribution to excess lifetimes over a range of thresholds starting from 105 years and give the maximum likelihood estimates in Table 1. The largest excess lifetime, for Emma Morano, who died aged 117 years in 2017, after c_2, is censored, and the estimated shape $\widehat{\xi}_u$ for threshold u is close to zero, although its variability is large for high u.

Table 2 gives the point estimates and 90% confidence intervals; the numbers of exceedances at low thresholds are appreciable, so higher-order correction have little impact on inferences for ι then but difference emerges as the threshold increases. The estimated upper bounds are large or even infinite.

To confirm our findings, we can estimate the distribution of the likelihood root for ι using the bootstrap (cf. Lee and Young (2005)). We did not consider this approach in the simulation study, as its good properties have been checked in other contexts and its calibration entails a costly double parametric bootstrap. The bth bootstrap likelihood root $R^{(b)}(\iota)$ is computed at each value of ι based on a sample simulated from a generalized Pareto distribution with parameters $(\widehat{\xi}_\iota, \iota)$. Figure 8 shows that the bootstrap p-value and the p-value obtained from the asymptotic χ^2_1 distribution of the profile likelihood ratio test agree up to Monte Carlo variability and suggests that this approach may be useful more widely in the context of extremal inference.

TABLE 2

Point estimates (90% confidence intervals) for the upper limit to lifetime ι (in years) based on the profile likelihood ratio statistic (middle) and the empirical covariance-based penalized profile likelihood (right) using exceedances of u for the Italian semisupercentenarian data. Values above 1000 *years are replaced by —*

u	profile	Modified profile
105	142.2 (127.1, 282.9)	142.1 (127.1, 287.7)
105.5	129.9 (121.6, 142.5)	127.1 (121.6, 142.7)
106	129.3 (123.0, 174.6)	131.8 (123.0, 176.1)
106.5	131.7 (124.2, —)	139.1 (124.2, —)
107	195.5 (128.1, —)	193.1 (128.3, —)
108	210.3 (123.7, —)	199.1 (124.0, —)
108.5	139.0 (120.6, —)	143.1 (120.8, —)

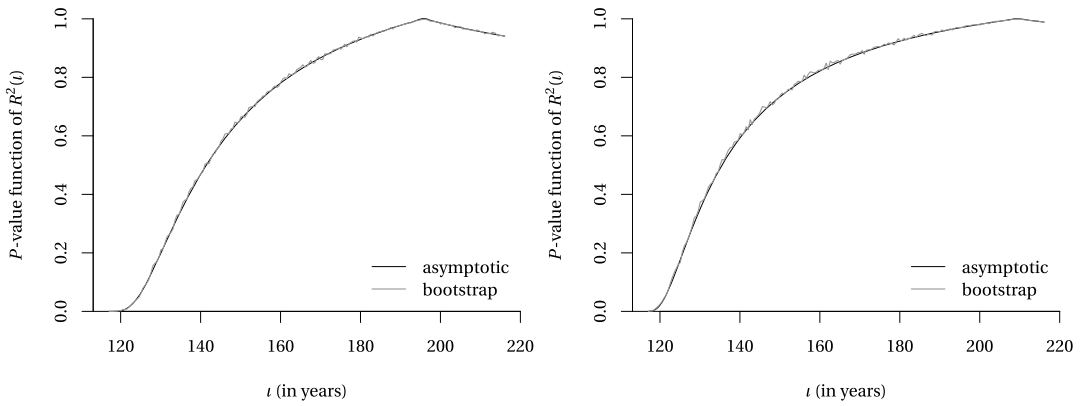

FIG. 8. *Italian semisupercentenarian data*: *p-value function for the profile likelihood ratio statistic* $R^2(\iota) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\xi}_\iota)\}$ *using excess lifetime above* $u = 107$ (*left*) *and* $u = 108$ (*right*) *years based on the asymptotic* χ^2_1 *distribution* (*black*) *and the bootstrap distribution* $\#\{b : R^{2(b)}(\iota) > R^2(\iota)\}/B$ *for bootstrap replications* $b = 1, \ldots, B$ (*grey*).

The apparent stability of the estimates and the large standard errors for $\widehat{\xi}_u$, seen in Table 1, do not allow us to rule out the exponential tail for thresholds $u > 107$ years, though the p-values for the four lowest thresholds 105–106.5 years, of 4.1%, 0.5%, 1.4%, and 6.0%, are smaller.

Under the exponential model the probability of surviving one additional year conditional on survival up to u years is $\exp(-1/\tau)$. Based on exceedances of $u = 110$ years, the model would yield an estimated probability of surviving an additional year of 0.476 with 95% confidence interval (0.416, 0.537): fewer than four in a thousand supercentenarians would be expected to live older than Emma Morano. These results are coherent with those of Rootzén and Zholud (2017), who analysed a smaller dataset on individuals who lived over 110 years.

5. Simulation study. The higher-order methods highlighted in the data illustrations typically lead to much wider confidence intervals for high quantiles. Despite their theoretically appealing properties, one might ask whether the additional effort is worth it, and, in particular, whether profile likelihood intervals have adequate coverages even though sample sizes for extremes, whether block maxima or threshold exceedances, are often small. We used Monte Carlo simulation to investigate small-sample inference for risk measures, based on the profile log-likelihood, the tangent exponential model approximation and Severini's corrections.

5.1. *General setup*. In a typical data analysis, one may attempt to predict the 100-year maximum temperature based on 20 years of daily records, where restricting attention to summer months yields around 90 observations per year. To mimic this scenario, we generated 1800 independent observations from a parametric model and targeted the expectation and median of the distribution of 9000-observation maximum from that same distribution with benchmarks computed using penultimate approximations.

The choice of block size or threshold involves a compromise between closeness of approximation (and thus reduced asymptotic bias) and small-sample effects. For larger block size/thresholds, the extreme-value approximation is, in principle, better, but estimation uncertainty is larger because of the smaller sample size. We divided the 1800 simulated values into blocks of sizes $m = 30, 45, 90$, and fitted the GEV distribution to the block maxima. We also fitted the GP distribution to the largest $n_u = 20, 40, 60$ order statistics of a sample of size 1800 from the GP distribution. We likewise generated data from six other distributions

mentioned in Section 2.3 and applied both block maximum and threshold methods to these data; see the Supplementary Material (Belzile and Davison (2022)).

For each sample we obtained four estimates and five sets of confidence limits for ψ, based on the Wald statistic; the likelihood root $R(\psi)$ and the modified likelihood root $R^\star(\psi)$ defined in equation (5) and equation (6); and the modified profile likelihoods (9) and (10). The Wald statistic was computed on the log scale and back-transformed, that is, with limits $\exp\{\log(\widehat{\psi}) \pm \Phi^{-1}(1-\alpha/2)\,\mathrm{se}(\widehat{\psi})/\widehat{\psi}\}$; the log transformation is intended to mitigate the poor properties of this statistic in highly asymmetric situations. For each target (return level, median, and expectation of the T-year maximum), distribution, and threshold or block size, we also calculated the relative bias of the point estimators as well as the overall coverage and the average widths of two-sided confidence intervals. The full results are in the Supplementary Material, and we summarize the main findings below, focusing on properties of one-sided confidence limits.

The maximum likelihood estimator of the shape parameters can occasionally be very large, leading to very wide confidence intervals. To avoid this unduly affecting the results, we use trimmed mean estimates for the relative width and the relative bias with 10% trimmed proportion in each tail.

5.2. *Summary of findings*. The reader is referred to the Supplementary Material for tables reporting the one-sided relative errors and widths of confidence intervals and the relative biases of point estimators.

Relative error of one-sided confidence intervals. Figures 9 and 10 display one-sided relative coverage errors for the expected N-observation maximum; similar results hold for the N-observation median and the N-observation return level. Despite the log-transformation, the Wald intervals fail to capture the positive skewness of the estimators of z_N, $\mathfrak{q}_{1/2}$ (not shown) and $\mathfrak{e}_N$ defined in Section 2.2. The one-sided relative coverage errors for the Wald statistics are so large that they fall outside the limits of Figures 9 and 10: for example, applying the block maximum method with 20 observations to samples from a GEV distribution, the Wald-based 99% confidence intervals contain the true value roughly 85% and 81% of the time when $\xi = 0.1$ and $\xi = -0.1$, respectively, but the 5% empirical error rate for the lower limit is 0%, indicating that the interval is too wide on the left and too short on the right.

If the data are generated from the generalized extreme value distribution, the empirical error rates for the TEM are closer to nominal, but no method is universally best. Perhaps unexpectedly, the penultimate effects are not really visible for the other distributions (see Supplementary Material). The profile and higher-order methods for block maxima seem impervious to the effects of extrapolation, and their coverage is excellent overall.

Figure 10 shows that the results based on threshold exceedances are more variable. The performance of Wald-based intervals remains calamitous: the empirical upper error rate for the nominal 5% limit is around 30% in all scenarios for the untransformed Wald statistic and improves only to 20–30% after transformation. With $k = 20$ observations (Figure 10), most higher-order methods overcover, even when the model is correctly specified. The TEM interval is shifted to the right, whereas Severini's corrections display higher empirical error in the lower tail. This breakdown of the TEM could be due to penultimate effects and small-sample bias, as it vanishes as the sample size grows; the TEM performs very well when $k = 60$. Two-sided profile likelihood intervals typically have good coverage, but their upper empirical error rates can be more than double the nominal values, as the intervals tend to lie too far to the left. Thus, the price paid for intervals with better coverage is increased uncertainty stemming from their greater width.

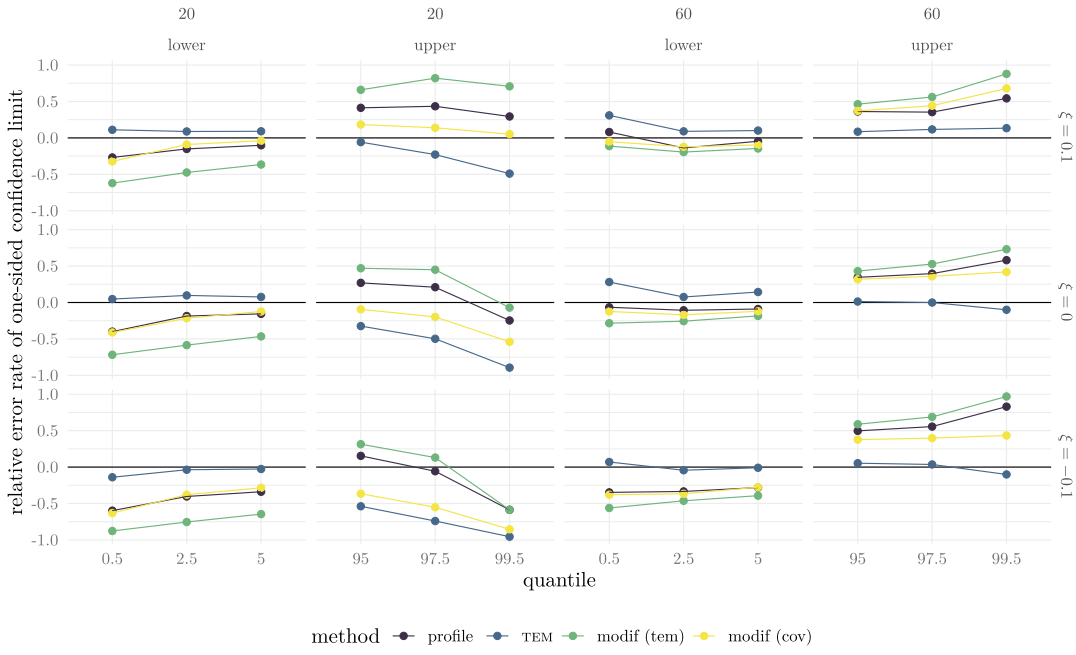

FIG. 9. *Relative coverage errors for one-sided lower and upper confidence limits with nominal error rates* 0.5%, 2.5%, *and* 5% *for the expected* 9000*-observation maximum, estimated using maxima of generalized extreme value samples of sizes* $n = 20$ (*left*) *and* $n = 60$ (*right*) *with shape parameter* $\xi = -0.1, 0, 0.1$ (*bottom to top*). *An ideal method would have zero relative error in both tails, whereas methods with relative error* ± 0.5 *have empirical error rates* 1.5 (+) *or* 0.5 (−) *times the nominal rate. The upper- and lower-tail errors for intervals whose relative errors have opposite signs will cancel to some extent when a two-sided interval is computed. If both upper- and lower-tail errors are positive, the corresponding two-sided intervals have empirical coverage that is too low, whereas negative upper- and lower-tail errors correspond to conservative two-sided confidence intervals.*

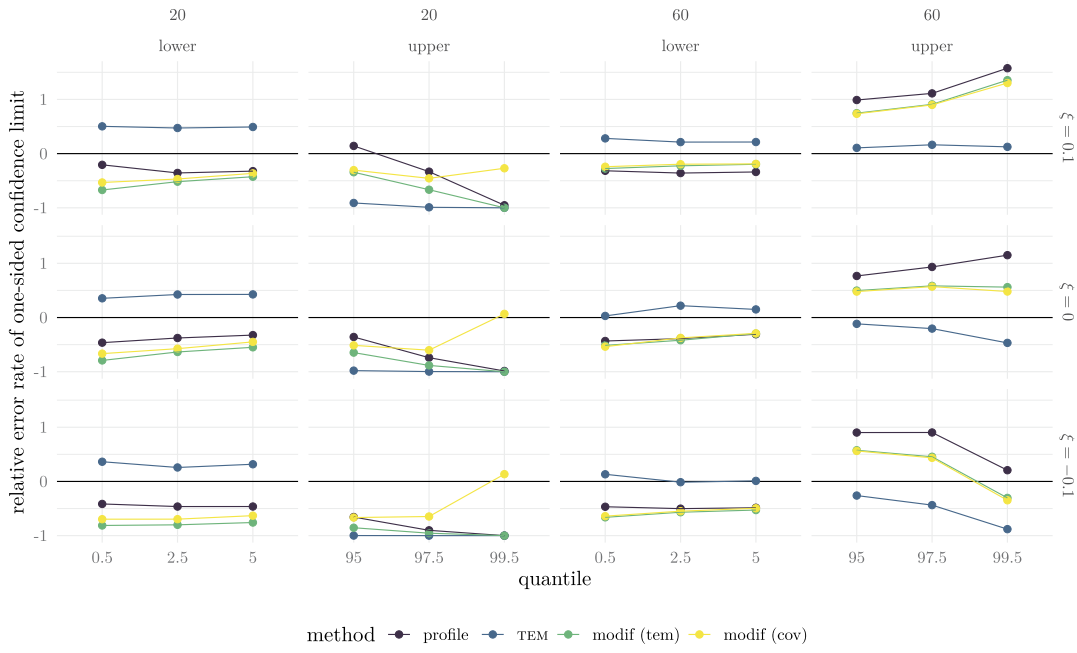

FIG. 10. *Relative coverage errors for one-sided lower and upper confidence limits for the expected* 9000*-observation maximum, estimated using generalized Pareto samples of sizes* $n = 20$ (*left*) *and* $n = 60$ (*right*) *with* $\xi = -0.1, 0, 0.1$ (*bottom to top*); *see the caption to Figure* 9 *for explanation.*

Width of confidence intervals. When $\xi > 0$, the expected N-observation maximum is larger than both the median of the N-observation maximum and the N-year return level, and its confidence intervals are the widest of those for all three risk measures, due to the extrapolation in the upper tail. The higher-order intervals, especially those based on $R^\star$, overcover slightly when the sample size is smaller and the blocks are larger, for example, for $m = 90$ with $k = 20$. The average widths of two-sided confidence intervals for the block maximum method with $m = 30$, $k = 60$ are comparable (not shown). For this setting the intervals based on $\ell_{\mathsf{m}}^{\mathrm{tem}}$ are the shortest among those implemented.

Higher-order methods for threshold exceedances give wider confidence intervals, often because they have better coverage in the upper tail: for example, the TEM confidence intervals are between 1.75 and two times wider than those based on the profile likelihood when $k = 20$ and about 1.25 times wider when $k = 60$.

Bias of point estimators of risk measures. When using threshold exceedances, maximum likelihood estimators of ξ are negatively biased for any sample size $k \leq 60$ (Figure 2) and risk estimators are likewise downwardly biased. For $k = 20$, the TEM point estimators, obtained by solving the equation $R^\star(\psi) = 0$, are positively biased, but they have the lowest bias of all point estimators considered when $k \geq 40$. The point estimators, derived using Severini's modified profile log-likelihoods, have lower bias than the maximum likelihood estimator.

5.3. *Practical guidelines*. Wald-based confidence intervals for the risk measures considered here should never be used; their coverage is appallingly low, even after transformation. For block maxima, profile likelihood-based confidence intervals have good two-sided coverage overall for the risk measures we considered, and there seems to be little gain in using higher-order methods: the discrepancy between the empirical error rates in the lower and upper tails seems to be due to the bias of the risk estimators themselves. For both types of data, the TEM-based estimator is systematically larger than the maximum likelihood estimator. For threshold exceedances, TEM-based confidence intervals have very good coverage, and the corresponding point estimators have smaller bias when the sample size is larger than around 50; while no higher-order method is always better, TEM-based intervals usually improve on the others. Severini's modified profile, based on empirical covariance, can be seen as a compromise; in theory, it is less accurate than the TEM by $\mathrm{O}(n^{-1/2})$ in both moderate and large deviation senses for continuous data, but it is more easily derived.

6. Discussion. Our data analyses illustrate the use of higher-order likelihood methods in risk analysis. Extreme value approximations are only valid asymptotically, so it is tempting to select large block sizes or high thresholds to reduce model misspecification. In this case samples are often small, so first-order methods may perform poorly and the higher-order methods may give better inferences. We have accounted for model misspecification in simulations for identically distributed data; our results show that, although the coverage of two-sided profile likelihood confidence intervals is adequate, they tend to be shifted to the left, and higher-order methods lead to much wider intervals. The improved properties of higher-order approximations have been established in regression and other contexts (e.g., Fraser, Wong and Wu (1999)), and there is no reason to suppose that they would be different here; the heuristic argument for why they work is purely likelihood-based, so the precise form of the underlying model is unimportant, provided the usual regularity conditions for likelihood inference are satisfied (Davison and Reid (2022)). In formal proofs of accuracy, the accumulation of information is usually made explicit by supposing that the sample size n is increasing, but the approximations can work well, even with n very small, so higher-order approximations for more general regression models, such as that of Wang and Tsai (2009), could be envisioned.

Comparisons with profile-based intervals can provide reassurance that standard methods are adequate, but the latter may fare poorly when the dimension of the nuisance vector $\boldsymbol{\lambda}$ is large, in which case the TEM may be more useful.

Very similar Bayesian approximations exist, though if prior information is limited, then it may be wise to choose a prior that will give these approximations good frequentist properties. Their derivations typically involve integrating a double Laplace approximation (e.g., Davison (2003), Section 11.3.1); see Fraser, Reid and Wu (1999), Tierney and Kadane (1986) and Tierney, Kass and Kadane (1989). Fraser (2011) discusses properties of these Bayesian procedures and observes that they disagree with higher-order frequentist procedures, except in location families. Fraser et al. (2016) propose a form of data-dependent Jeffreys prior on the contour of the profile likelihood that leads to third-order reproducibility, but this requires a different prior for every risk measure and so is incoherent. Generalized additive models and penalized likelihood estimators (e.g., Mhalla, de Carvalho and Chavez-Demoulin (2019)) could be incorporated into this framework.

A large body of work in extremes concerns semiparametric estimation of a positive shape parameter ξ using variants of the Hill (1975) estimator, sometimes coupled with the Weissman (1978) quantile estimator. Recent proposals address the asymptotic bias of the resulting estimators under second-order regular variation (cf. de Haan and Ferreira (2006), Section 2.3) which involves the estimation of additional parameters. Examples include Gomes and Pestana (2007), who attempt to eliminate the asymptotic bias of log-quantile estimators, and Figueiredo et al. (2012) who consider peaks over random threshold Hill-type quantile estimators. In many cases, such estimators are asymptotically normal and uncertainty is expressed through Wald-type confidence intervals (e.g., de Haan and Ferreira (2006), Section 4.3); see Buitendag, Beirlant and de Wet (2020) for a recent alternative. As the inclusion of further parameters makes extremal inferences less precise and symmetric confidence intervals for risk measures have dire properties, we prefer to investigate the performance of likelihood-based approaches under misspecification rather than to consider these semiparametric methods. Likelihood-based methods readily apply to settings in which there is nonstationarity, selection mechanisms, or other complications, whereas semiparametric approaches typically require fresh theoretical work in each case.

The methods described above may usefully be generalized to risk measures for multivariate extremes, since, as the dimension grows, the performance of first-order methods can be expected to deteriorate, due to the increased numbers of nuisance parameters. Multivariate extreme value distributions must obey a moment constraint, and conclusions may be expected to depend on the chosen model more strongly than in the univariate case. Using empirical likelihood, Einmahl and Segers (2009) fit such a model, and de Carvalho and Davison (2014) extend this to incorporate covariates; as the properties of empirical likelihood closely mimic those of regular likelihood, adapting the TEM to this setting could be an avenue for future work.

APPENDIX A: INFORMATION MATRIX FOR THE r LARGEST ORDER STATISTICS

To derive the information matrix for the r-largest likelihood given in Equation (15), we note that the marginal density of Y_r is

$$f_{Y_r}(y_r;\mu,\sigma,\xi)=\frac{1}{(r-1)!\sigma}\left(1+\xi\frac{y_r-\mu}{\sigma}\right)_+^{-r/\xi-1}\exp\left\{-\left(1+\xi\frac{y_r-\mu}{\sigma}\right)_+^{-1/\xi}\right\},$$

so the joint density of $Y_1,\ldots,Y_r$ may be written as

$$\frac{1}{(r-1)!\sigma}\left(1+\xi\frac{y_r-\mu}{\sigma}\right)_+^{-r/\xi-1}\exp\left\{-\left(1+\xi\frac{y_r-\mu}{\sigma}\right)_+^{-1/\xi}\right\}$$

$$\times (r-1)! \prod_{j=1}^{r-1} \frac{1}{\sigma} \frac{(1+\xi\frac{y_j-\mu}{\sigma})_+^{-1/\xi-1}}{(1+\xi\frac{y_r-\mu}{\sigma})_+^{-1/\xi}};$$

that is, we write the joint density as the product of the density of Y_r and the joint conditional density of $Y_1, \ldots, Y_{r-1}$ conditional on $Y_r = y_r$. This conditional density equals that of the order statistics of $r-1$ independent variables with generalized Pareto density

$$h(\boldsymbol{y}_{-r} - y_r; \tau, \xi) = \prod_{j=1}^{r-1} \frac{1}{\tau}\left(1+\xi\frac{y_j - y_r}{\tau}\right)_+^{-1/\xi-1},$$

where $\tau = \sigma + \xi(y_r - \mu)$. Thus, the overall log-likelihood is

$$\ell(\mu, \sigma, \xi; y_j, y_r) \equiv \log\{f_{Y_r}(y_r; \mu, \sigma, \xi)\} + \sum_{j=1}^{r-1} \log\{h(y_j - y_r; \tau, \xi)\},$$

where $y_1, \ldots, y_{r-1}$ represent the observed values of a random sample of generalized Pareto variables. We may thus write the joint density of the r-largest order statistics as the product of the density of Y_r and the joint conditional density of $Y_1, \ldots, Y_{r-1}$, conditional on $Y_r = y_r$. To obtain the observed information, we first calculate the Hessian matrix of $-\ell$, then condition on $Y_r = y_r$ and take expectations over $X_j = Y_j - y_r$. It remains to write $\tau = \sigma + \xi(y_r - \mu)$ and integrate over Y_r. The matrices themselves can be found in the Supplementary Material, and a numerical implementation is available via the function `rlarg.infomat` in the R package `mev`.

APPENDIX B: THE TEM FOR THE POISSON PROCESS

Suppose we observe events of an inhomogeneous Poisson process $\mathcal{P}$ with intensity $\dot{\nu}(x; \boldsymbol{\theta})$ for $x \in \mathcal{X}$, where $\mathcal{X}$ is partitioned into subsets $\mathcal{X}_1, \ldots, \mathcal{X}_K$. Let $N(\mathcal{A})$ denote the number of events of $\mathcal{P}$ in a measurable set $\mathcal{A} \subset \mathcal{X}$, let $N_k = N(\mathcal{X}_k)$, and suppose that $N(\mathcal{X})$ has finite expectation

$$\nu(\mathcal{X}; \boldsymbol{\theta}) = \int_{\mathcal{X}} \dot{\nu}(x; \boldsymbol{\theta})\, \mathrm{d}x.$$

If $\dot{\nu}$ is constant on each of the $\mathcal{X}_k$, then the log-likelihood is that of the independent Poisson variables $N_1, \ldots, N_K$,

$$\begin{aligned}\sum_{k=1}^{K} n_k \log\{|\mathcal{X}_k|\dot{\nu}(x_k; \boldsymbol{\theta})\} - |\mathcal{X}_k|\dot{\nu}(x_k; \boldsymbol{\theta}) &\equiv \sum_{k=1}^{K}\{n_k \log \dot{\nu}(x_k; \boldsymbol{\theta}) - |\mathcal{X}_k|\dot{\nu}(x_k; \boldsymbol{\theta})\} \\ &= \sum_{k=1}^{K} \ell_k(\boldsymbol{\theta}),\end{aligned}$$

say, where n_k is the realised value of N_k and $x_k \in \mathcal{X}_k$. The terms $n_k \log|\mathcal{X}_k|$, dropped at the $\equiv$ sign, do not depend on $\boldsymbol{\theta}$; retaining them leads to an affine transformation of $\varphi(\boldsymbol{\theta})$ and makes no difference to inferences. The arguments in Davison, Fraser and Reid (2006) imply that second-order inference is obtained on using

$$\varphi(\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbf{V}_k \frac{\partial \ell_k(\boldsymbol{\theta})}{\partial n_k},$$

where

$$\mathbf{V}_k = \frac{\partial \mathsf{E}(N_k; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \frac{\partial |\mathcal{X}_k|\dot{\nu}(x_k; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}},$$

and this yields

$$\varphi(\boldsymbol{\theta}) = \sum_{k=1}^{K} |\mathcal{X}_k| \frac{\partial \dot{\nu}(x_k;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \log \dot{\nu}(x_k;\boldsymbol{\theta}) = \int_{\mathcal{X}} \frac{\partial \dot{\nu}(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \log \dot{\nu}(x;\boldsymbol{\theta})\,\mathrm{d}x. \tag{17}$$

This integral does not depend on the partition of $\mathcal{X}$ and so must be the limit as $K \to \infty$.

Since Equation (17) is intractable in general, we employ the numerical approximation

$$\sum_{j=1}^{n} \frac{\partial \dot{\nu}(x_j;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \times \frac{1}{\dot{\nu}(x_j;\boldsymbol{\theta})} \log \dot{\nu}(x_j;\boldsymbol{\theta}), \tag{18}$$

based on events $x_1, \ldots, x_n \in \mathcal{X}$, which differs from $\varphi(\boldsymbol{\theta})$ by a term of order $\nu(\mathcal{X};\boldsymbol{\theta})^{1/2}$ and, therefore, gives the same order of error. To check this, note that, conditional on $N(\mathcal{X}) = n$, the x_j are independent and identically distributed on $\mathcal{X}$ with density $\dot{\nu}(x;\boldsymbol{\theta})/\nu(\mathcal{X};\boldsymbol{\theta})$. Thus, the expectation of (18), conditional on $N(\mathcal{X}) = n$, is

$$n \int_{\mathcal{X}} \frac{\partial \dot{\nu}(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \times \frac{1}{\dot{\nu}(x;\boldsymbol{\theta})} \log \dot{\nu}(x;\boldsymbol{\theta}) \times \frac{\dot{\nu}(x;\boldsymbol{\theta})}{\nu(\mathcal{X};\boldsymbol{\theta})}\,\mathrm{d}x = \frac{n}{\nu(\mathcal{X};\boldsymbol{\theta})}\varphi(\boldsymbol{\theta}),$$

and, as $N(\mathcal{X})$ has expectation $\nu(\mathcal{X};\boldsymbol{\theta})$, the expectation of (18) is $\varphi(\boldsymbol{\theta})$, as required. One can verify that (18) has variance of order $\nu(\mathcal{X};\boldsymbol{\theta})$ under mild conditions on the integrand.

APPENDIX C: THE TEM FOR THE MAIQUETÍA EXAMPLE

We illustrate the derivation of the tangent exponential model from Section 4.1, with threshold exceedances $Y_1, \ldots, Y_n, Y_t$ independent of the yearly maximum for $Z_1, \ldots, Z_{10}$: the contribution of each to the TEM is additive with $\boldsymbol{\varphi}(\boldsymbol{\theta}) = \boldsymbol{\varphi}^{\boldsymbol{Z}}(\boldsymbol{\theta}) + \boldsymbol{\varphi}^{\boldsymbol{Y}}(\boldsymbol{\theta})$ and, similarly, $\boldsymbol{\varphi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \boldsymbol{\varphi}^{\boldsymbol{Z}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} + \boldsymbol{\varphi}^{\boldsymbol{Y}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. We suppose that the yearly maxima are identically distributed with $Z_i \sim \mathsf{GEV}(\mu, \sigma, \xi)$. The model has parameters $\psi = \mathfrak{q}_p$ and $\boldsymbol{\lambda} = (\sigma, \xi)$. The generalized extreme value distribution is a location-scale family, so the ith row of the 10×3 matrix of sufficient directions has components $\mathrm{V}_{z_i,\mu} = -1$ and $\mathrm{V}_{z_i,\sigma} = -(z_i^\circ - \widehat{\mathfrak{q}}_p)/\widehat{\sigma}$, while

$$\mathrm{V}_{z_i,\widehat{\xi}} = \frac{\widehat{\sigma}\widehat{c}_\xi \log(\widehat{c}_\xi) - \widehat{\xi}(z_i^\circ - \mathfrak{q}_p) - \{\widehat{\xi}(z_i^\circ - \mathfrak{q}_p) + \widehat{c}_\xi\}\log(\widehat{c}_\xi - \widehat{\xi}\frac{z_i^\circ - \widehat{\mathfrak{q}}_p}{\widehat{\sigma}})}{\widehat{\xi}^2}.$$

Let $t(z_i) = 1 + \xi(z_i - \mu)/\sigma$, where μ is an implicit function of $(\mathfrak{q}_p, \sigma, \xi)$; the canonical parameters are obtained by multiplying the vector of sample space derivatives by the matrix of sufficient directions,

$$\boldsymbol{\varphi}^{\boldsymbol{Z}}(\boldsymbol{\theta}) = \sum_{i=1}^{10} \mathbf{V}_{z_i}^\top \times \frac{t(z_i)^{-\frac{1}{\xi}-1} + (1+\xi)t(z_i)^{-1}}{\sigma} = \sum_{i=1}^{10} \mathbf{V}_{z_i}^\top \times g(z_i),$$

say, and the rows of the matrix of mixed derivatives are

$$\boldsymbol{\varphi}^{\boldsymbol{Z}}_{\mathfrak{q}_p}(\boldsymbol{\theta}) = \sum_{i=1}^{10} \mathbf{V}_{z_i}^\top \times \frac{(1+\xi)t(z_i)^{-1/\xi-2} + \xi(1+\xi)t(z_i)^{-2}}{\sigma^2},$$

$$\boldsymbol{\varphi}^{\boldsymbol{Z}}_{\sigma}(\boldsymbol{\theta}) = \sum_{i=1}^{10} \mathbf{V}_{z_i}^\top \times \left\{ \frac{g(z_i)}{\sigma} - \frac{(1+1/\xi)t(z_i)^{-1/\xi-2} + (1+\xi)t(z_i)^{-2}}{\sigma} \frac{\partial t(z_i)}{\partial \sigma} \right\},$$

etc.

The arguments in Section 4.1 show that the distribution function of the n first threshold exceedances below s, $\Pr(Y_i + u \leq y) = \{\Lambda(u) - \Lambda(y)\}/\{\Lambda(u) - \Lambda(s)\}$, while the largest observation $Y_t \equiv Y_{n+1}$ is left-truncated at s and, by virtue of the threshold stability property, $Y_t - s$ follows a generalized Pareto distribution with scale $\sigma + \xi(s - \mu)$ and shape ξ. We can thus derive the sufficient directions for the threshold exceedances as usual, with

$$\mathbf{V}_{y_{n+1},\mathfrak{q}_p} = -\frac{\widehat{\xi}(y_t^\circ - s)}{\widehat{\sigma}\widehat{c}_\xi + \widehat{\xi}(s - \widehat{\mathfrak{q}}_p)}, \qquad \mathbf{V}_{y_{n+1},\sigma} = \frac{\widehat{c}_\xi(y_t^\circ - s)}{\widehat{\sigma}\widehat{c}_\xi + \widehat{\xi}(s - \widehat{\mathfrak{q}}_p)},$$

and so forth. Truncation only affects the sufficient directions and, for all of the $n + 1$ exceedances, the sample space and mixed derivatives are of the form

$$\boldsymbol{\varphi}^{\boldsymbol{Y}}(\boldsymbol{\theta}) = \sum_{i=1}^{n+1} \mathbf{V}_{y_i}^\top \times \frac{-(1+\xi)}{\sigma c_\xi + \xi(y_i^\circ - \mathfrak{q}_p)},$$

$$\boldsymbol{\varphi}^{\boldsymbol{Y}}_{\mathfrak{q}_p}(\boldsymbol{\theta}) = \sum_{i=1}^{n+1} \mathbf{V}_{y_i}^\top \times \frac{-(1+\xi)\xi}{\{\sigma c_\xi + \xi(y_i^\circ - \mathfrak{q}_p)\}^2},$$

$$\boldsymbol{\varphi}^{\boldsymbol{Y}}_{\sigma}(\boldsymbol{\theta}) = \sum_{i=1}^{n+1} \mathbf{V}_{y_i}^\top \times \frac{(1+\xi)c_\xi}{\{\sigma c_\xi + \xi(y_i^\circ - \mathfrak{q}_p)\}^2},$$

$$\boldsymbol{\varphi}^{\boldsymbol{Y}}_{\xi}(\boldsymbol{\theta}) = \sum_{i=1}^{n+1} \mathbf{V}_{y_i}^\top \times \frac{\sigma c_\xi\{\log(c_\xi) - \xi\} + \xi(y_i^\circ - \mathfrak{q}_p)}{\xi\{\sigma c_\xi + \xi(y_i^\circ - \mathfrak{q}_p)\}^2}.$$

Acknowledgments. We thank the reviewers for particularly useful comments. This work was performed using the R programming language (R Core Team (2021)) with formulae derived through SageMath (The Sage Developers (2021)).

Funding. This research was financially supported by the Natural Sciences and Engineering Research Council of Canada and the Swiss National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplement to "Improved inference on risk measures for univariate extremes" (DOI: 10.1214/21-AOAS1555SUPPA; .pdf). Expressions for the entries of the information matrices of the r-largest order statistics of a generalized extreme value distribution. Description of the simulation study infrastructure and set-up. Additional results from the simulation study, including relative coverage and width of confidence intervals and relative bias for the block maximum and the peaks-over-threshold methods.

Code and data (DOI: 10.1214/21-AOAS1555SUPPB; .zip). The compressed archive `AOAS1555-sm.zip` contains the R code used to generate all the results presented in the paper. It is also available from https://github.com/lbelzile/hoa-extremes.

REFERENCES

BARBI, E., LAGONA, F., MARSILI, M., VAUPEL, J. W. and WACHTER, K. W. (2018). The plateau of human mortality: Demography of longevity pioneers. *Science* **360** 1459–1461. MR3792342 https://doi.org/10.1126/science.aat3119

BARLOW, A. M., SHERLOCK, C. and TAWN, J. (2020). Inference for extreme values under threshold-based stopping rules. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 765–789. MR4133146

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics. Monographs on Statistics and Applied Probability* **52**. CRC Press, London. MR1317097 https://doi.org/10.1007/978-1-4899-3210-5

BELZILE, L. R. (2019). Contributions to likelihood-based modelling of extreme values. Ph.D. thesis. EPFL, Lausanne.

BELZILE, L. R and DAVISON, A. C (2022). Supplement to "Improved inference on risk measures for univariate extremes." https://doi.org/10.1214/21-AOAS1555SUPPA, https://doi.org/10.1214/21-AOAS1555SUPPB

BELZILE, L. R., DAVISON, A. C., ROOTZÉN, H. and ZHOLUD, D. (2021). Human mortality at extreme age. *R. Soc. Open Sci.* **8** 202097. https://doi.org/10.1098/rsos.202097

BELZILE, L. R., DAVISON, A. C., GAMPE, J., ROOTZÉN, H. and ZHOLUD, D. (2022). Is there a cap on longevity? A statistical review. *Annu. Rev. Stat. Appl.* **9**, in press. https://doi.org/10.1146/annurev-statistics-040120-025426

BRAZZALE, A. R., DAVISON, A. C. and REID, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **23**. Cambridge Univ. Press, Cambridge. MR2342742 https://doi.org/10.1017/CBO9780511611131

BÜCHER, A. and SEGERS, J. (2017). On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes* **20** 839–872. MR3737387 https://doi.org/10.1007/s10687-017-0292-6

BUITENDAG, S., BEIRLANT, J. and DE WET, T. (2020). Confidence intervals for extreme Pareto-type quantiles. *Scand. J. Stat.* **47** 36–55. MR4075228 https://doi.org/10.1111/sjos.12396

COLES, S. and PERICCHI, L. (2003). Anticipating catastrophes through extreme value modelling. *J. Roy. Statist. Soc. Ser. C* **52** 405–416. MR2012566 https://doi.org/10.1111/1467-9876.00413

COLES, S., PERICCHI, L. R. and SISSON, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *J. Hydrol.* **273** 35–50.

COX, D. R., ISHAM, V. S. and NORTHROP, P. J. (2002). Floods: Some probabilistic and statistical approaches. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **360** 1389–1408.

COX, D. R. and SNELL, E. J. (1968). A general definition of residuals (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 248–275. MR0237052

DAVIS, R. A. and MIKOSCH, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli* **15** 977–1009. MR2597580 https://doi.org/10.3150/09-BEJ213

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73** 323–332. MR0855892 https://doi.org/10.1093/biomet/73.2.323

DAVISON, A. C. (2003). *Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913 https://doi.org/10.1017/CBO9780511815850

DAVISON, A. C., FRASER, D. A. S. and REID, N. (2006). Improved likelihood inference for discrete data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 495–508. MR2278337 https://doi.org/10.1111/j.1467-9868.2006.00548.x

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673 https://doi.org/10.1017/CBO9780511802843

DAVISON, A. C. and REID, N. (2022). The tangent exponential model. In *Handbook of Bayesian, Fiducial and Frequentist Inference* (J. O. Berger, X. L. Meng, N. Reid and M. Xie, eds.) CRC Press/CRC, Boca Raton, FL.

DE CARVALHO, M. and DAVISON, A. C. (2014). Spectral density ratio models for multivariate extremes. *J. Amer. Statist. Assoc.* **109** 764–776. MR3223748 https://doi.org/10.1080/01621459.2013.872651

DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction. Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2234156 https://doi.org/10.1007/0-387-34471-3

DOMBRY, C. and FERREIRA, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli* **25** 1690–1723. MR3961227 https://doi.org/10.3150/18-BEJ1032

EINMAHL, J. J., EINMAHL, J. H. J. and DE HAAN, L. (2019). Limits to human life span through extreme value theory. *J. Amer. Statist. Assoc.* **114** 1075–1080. MR4011759 https://doi.org/10.1080/01621459.2018.1537912

EINMAHL, J. H. J. and SEGERS, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Statist.* **37** 2953–2989. MR2541452 https://doi.org/10.1214/08-AOS677

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events—for Insurance and Finance. Applications of Mathematics* **33**. Springer, Berlin. MR1458613 https://doi.org/10.1007/978-3-642-33483-2

FASIOLO, M., WOOD, S. N., ZAFFRAN, M., NEDELLEC, R. and GOUDE, Y. (2021). Fast calibrated additive quantile regression. *J. Amer. Statist. Assoc.* **116** 1402–1412. MR4309281 https://doi.org/10.1080/01621459.2020.1725521

FIGUEIREDO, F., GOMES, M. I., HENRIQUES-RODRIGUES, L. and MIRANDA, M. C. (2012). A computational study of a quasi-PORT methodology for VaR based on second-order reduced-bias estimation. *J. Stat. Comput. Simul.* **82** 587–602. MR2908951 https://doi.org/10.1080/00949655.2010.547196

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. MR1225212 https://doi.org/10.1093/biomet/80.1.27

FISHER, R. A. and TIPPETT, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math. Proc. Cambridge Philos. Soc.* **24** 180–190.
FRASER, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26** 299–316. MR2918001 https://doi.org/10.1214/11-STS352
FRASER, D. A. S., REID, N. and WU, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86** 249–264. MR1705367 https://doi.org/10.1093/biomet/86.2.249
FRASER, D. A. S., WONG, A. and WU, J. (1999). Regression analysis, nonlinear or nonnormal: Simple and accurate p values from likelihood analysis. *J. Amer. Statist. Assoc.* **94** 1286–1295. MR1731490 https://doi.org/10.2307/2669942
FRASER, D. A. S., BÉDARD, M., WONG, A., LIN, W. and FRASER, A. M. (2016). Bayes, reproducibility and the quest for truth. *Statist. Sci.* **31** 578–590. MR3598740 https://doi.org/10.1214/16-STS573
GILES, D. E., FENG, H. and GODWIN, R. T. (2016). Bias-corrected maximum likelihood estimation of the parameters of the generalized Pareto distribution. *Comm. Statist. Theory Methods* **45** 2465–2483. MR3480663 https://doi.org/10.1080/03610926.2014.887104
GNEDENKO, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. of Math.* (2) **44** 423–453. MR0008655 https://doi.org/10.2307/1968974
GOMES, M. I. and PESTANA, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *J. Amer. Statist. Assoc.* **102** 280–292. MR2345543 https://doi.org/10.1198/016214506000000799
HANAYAMA, N. and SIBUYA, M. (2016). Estimating the upper limit of lifetime probability distribution, based on data of Japanese centenarians. *J. Gerontol., Ser. A* **71** 1014–1021.
HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204
HOSKING, J. R. M. and WALLIS, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* **29** 339–349. MR0906643 https://doi.org/10.2307/1269343
KENNE PAGUI, E. C., SALVAN, A. and SARTORI, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* **104** 923–938. MR3737312 https://doi.org/10.1093/biomet/asx046
LEE, S. M. S. and YOUNG, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Statist. Probab. Lett.* **71** 143–153. MR2126770 https://doi.org/10.1016/j.spl.2004.10.026
MHALLA, L., DE CARVALHO, M. and CHAVEZ-DEMOULIN, V. (2019). Regression-type models for extremal dependence. *Scand. J. Stat.* **46** 1141–1167. MR4033807 https://doi.org/10.1111/sjos.12388
PICKANDS, J. III (1986). The continuous and differentiable domains of attraction of the extreme value distributions. *Ann. Probab.* **14** 996–1004. MR0841599
PIRAZZOLI, P. A. (1982). Maree estreme a Venezia (periodo 1872–1981). *Acqua Aria* **10** 1023–1039.
PIRES, J. F., CYSNEIROS, A. H. M. A. and CRIBARI-NETO, F. (2018). Improved inference for the generalized Pareto distribution. *Braz. J. Probab. Stat.* **32** 69–85. MR3770864 https://doi.org/10.1214/16-BJPS332
R CORE TEAM (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
ROODMAN, D. (2018). Bias and size corrections in extreme value modeling. *Comm. Statist. Theory Methods* **47** 3377–3391. MR3803408 https://doi.org/10.1080/03610926.2017.1353630
ROOTZÉN, H. and ZHOLUD, D. (2017). Human life is unlimited—but short (with discussion). *Extremes* **20** 713–728. MR3737382 https://doi.org/10.1007/s10687-017-0305-5
SEVERINI, T. A. (2000). *Likelihood Methods in Statistics. Oxford Statistical Science Series* **22**. Oxford Univ. Press, Oxford. MR1854870
SKOVGAARD, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2** 145–165. MR1410135 https://doi.org/10.2307/3318548
SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90. MR0790201 https://doi.org/10.1093/biomet/72.1.67
SMITH, R. L. (1986). Extreme value theory based on the r largest annual events. *J. Hydrol.* **86** 27–43.
SMITH, R. L. (1987). Approximations in extreme value theory. Technical Report 205, Center for Stochastic Processes, University of North Carolina Chapel Hill.
THE SAGE DEVELOPERS (2021). SageMath, the Sage Mathematics Software System (Version 9.3).
TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567
TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika* **76** 425–433 (correction: **78**, 233–234). MR1040637 https://doi.org/10.1093/biomet/76.3.425
WADSWORTH, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics* **58** 116–126. MR3463162 https://doi.org/10.1080/00401706.2014.998345
WANG, H. and TSAI, C.-L. (2009). Tail index regression. *J. Amer. Statist. Assoc.* **104** 1233–1240. MR2750246 https://doi.org/10.1198/jasa.2009.tm08458

WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73** 812–815. MR0521329

WIECZOREK, G. F., LARSEN, M. C., EATON, L. S., MORGAN, B. A. and BLAIR, J. L. (2001). Debris-flow and flooding hazards associated with the December 1999 storm in coastal Venezuela and strategies for mitigation Technical Report No. 01-0144 U.S. Geological Survey.