The Annals of Statistics
2002, Vol. 30, No. 3, 875–895

EFFECT OF EXTRAPOLATION ON COVERAGE ACCURACY OF PREDICTION INTERVALS COMPUTED FROM PARETO-TYPE DATA

BY PETER HALL, LIANG PENG AND NADER TAJVIDI[1]

Australian National University, Australian National University and Lund Institute of Technology

A feature that distinguishes extreme-value contexts from more conventional statistical problems is that in the former we often wish to make predictions well beyond the range of the data. For example, one might have a 10-year sequence of observations of a phenomenon, and wish to make forecasts for the next 20 to 30 years. It is generally unclear how such long ranges of extrapolation affect prediction. In the present paper, and for extremes from a distribution with regularly varying tails at infinity, we address this problem. We approach it in two ways: first, from the viewpoint of predictive inference under a model that is admittedly only approximate, and where the errors of greatest concern are caused by the interaction of long-range extrapolation with model misspecification; second, where the model is accurate but errors arise from a combination of extrapolation and the fact that the method is only approximate. In both settings we show that, in a way which can be defined theoretically and confirmed numerically, one can make predictions exponentially far into the future without committing serious errors.

1. Introduction. Suppose a sequence of events, each with a numerical "strength," is observed over time. For example, the events might be windstorms, in which case the strength of an event could be the total cost to an insurance company of a storm. In such cases, attention would usually be confined to extreme events, and the very existence of a storm would be defined as an exceedence of a threshold. Given data on the strengths of previous events, we wish to construct prediction intervals to forecast future events.

In extreme-value contexts one often wishes to make predictions well beyond the range of the data. For example, it may be necessary to make predictions about the strengths of events over the next 25 years, based on data from the last 10 years. It is important to know how robust our prediction intervals are to this sort of extrapolation, assuming that the model persists over the time period. In mathematical terms we wish to know whether we can make predictions polynomially far into the future, or even exponentially far, from a given amount of

Received April 1999; revised September 2001.

[1]Research supported by the Swedish Research Council and Foundation for International Cooperation in Research and Higher Education.

AMS 2000 subject classifications. Primary 62G30; secondary 62G20.

Key words and phrases. Bootstrap, calibration, coverage accuracy, domain of attraction, exceedence, extreme value, generalized Pareto distribution, peaks over threshold.

data. That is, given a sample of n data values, can we provide accurate prediction bounds for extrema of the next $O(n^C)$ data, or even the next $O\{\exp(n^C)\}$ data, for some $C > 0$? In more practical terms we wish to know how such theoretical results translate into numerical accuracy of prediction bounds.

In the present paper, and in the context of extreme-value data with Pareto-type distributions, we address these questions. We take two viewpoints. First, we suppose the data are only approximately Pareto-distributed (e.g., that they are in the same extreme-value domain of attraction as the Pareto). There we assess the influence of departures from the Pareto model on long-range forecasts. Second, we assume the data are exactly distributed according to the generalized Pareto (GPD) model and assess the accuracy of long-range forecasts in this case.

Our second set of results complements the first by showing what is possible when extrema are accurately modelled, and bootstrap methods are used to correct for statistical errors, rather than model-misspecification errors. When the GPD model is valid, bootstrap calibration is one way of correcting for the inaccuracy of naive prediction, but it still does not give perfect accuracy. Just as in the case of unmodelled departure from the standard Pareto model, there are prediction errors, and these are exacerbated by long-range forecasting.

The conclusions of our analysis are broadly similar in both contexts: extrapolation can be made exponentially, rather than just polynomially, far into the future before serious errors due specifically to extrapolation are introduced. Expressed in theoretical terms, if the error for one-step-ahead prediction, based on a sample of size n, is $\delta = \delta(n)$, then the error for m-step-ahead prediction will be of order $\Delta = (\log m)^a \delta$ for some $a > 0$. Usually δ is polynomially small, that is, of order n^{-b} for some $b > 0$. In such cases m can be exponentially large, as a function of n, before extrapolation causes significant problems.

The value of a in the formula for Δ depends on context. For example, in the case of fitting a Pareto model to data that are only approximately Pareto-distributed, where the error δ arises through model misspecification, we generally have $a = 1$. This problem is explored in Section 2, where we also address bootstrap calibration approaches which help to reduce the coverage error of prediction intervals. In the case of constructing simple prediction intervals for GPD-distributed data, where δ is a consequence of inaccuracies of the method rather than of the model, the value of a will generally be larger for methods that are more accurate. This is not a problem, of course, since δ is smaller for more accurate methods, and its reduced size more than compensates for the larger value of the logarithmic factor. These issues are taken up in Section 3, where again we treat refinements based on bootstrap calibration.

For both problems—that is, prediction using an approximate model, and prediction using the correct model but an approximate method—we report numerical results which corroborate the conclusions of our mathematical analysis. We also discuss "interactions" between the problems, where bootstrap calibration is used in the presence of model misspecification. Technical arguments behind our

results are given in Section 4. For the sake of simplicity and brevity we confine attention to extremes in the upper, infinite tail of a distribution, where the tail is regularly varying at infinity. Of course, we could similarly treat extremes in a lower tail at the origin.

Applications of the Pareto and generalized Pareto models include the following: the early, relatively qualitative work of Zipf (1941, 1949); applications to hydrology, by Davison (1984), Smith (1984), Hosking and Wallace (1987), Davison and Smith (1990) and Moharram, Gosain and Kapoor (1993); applications to analysis of ozone levels, by Smith (1989); applications in the insurance industry, by Rytgaard (1990) and Rootzén and Tajvidi (1997); and applications to the study of fiber strength, by Grimshaw (1993). Properties of estimators of Pareto parameters have been studied by, in addition to the aforementioned authors, Davis and Resnick (1984), Csörgő, Deheuvels and Mason (1985), Smith (1985), Leadbetter (1991) and Rosbjerg, Madsen and Rasmussen (1992). Bootstrap prediction intervals in more conventional settings, not involving extensive extrapolation, have been studied by Stine (1985), Bai and Olshen (1988), Bai, Bickel and Olshen (1990) and Beran (1990, 1992).

2. Semiparametric prediction based on a Pareto model.

2.1. *Approximate models for the tail of a distribution.* We begin by describing tail properties of distributions conditional on exceedence of a high threshold, and then we discuss a model motivated by these properties. Let Y denote a random variable the distribution of which satisfies

$$P(Y > y) = a_1 y^{-\beta} + a_2 y^{-\beta-\gamma} + o(y^{-\beta-\gamma}) \tag{2.1}$$

as $y \to \infty$, where $a_1, \beta, \gamma > 0$ and $a_2 \in (-\infty, \infty)$. Then, for $t \geq 1$,

$$P(Y > y_0 t | Y > y_0) = A_1 t^{-\beta} + A_2 t^{-\beta-\gamma} + o(|A_2| t^{-\beta-\gamma})$$

as $y_0 \to \infty$, where

$$A_1 = 1 - \frac{a_2 y_0^{-\gamma}}{a_1 + a_2 y_0^{-\gamma} + o(y_0^{-\gamma})}, \qquad A_2 = \frac{a_2 y_0^{-\gamma}}{a_1 + a_2 y_0^{-\gamma} + o(y_0^{-\gamma})}.$$

Therefore, if $\mathcal{X} = \{X_1, \ldots, X_n\}$ denotes the n values of a random sample $\{Y_1, \ldots, Y_N\}$ that exceed a given high threshold y_0, and if we condition on n, then without loss of generality we have drawn $\mathcal{X}$ from a distribution which depends on n and for which $P(X_i > 1) = 1$ and

$$P(X_i > x) = (1 - \delta) x^{-\beta} + \delta x^{-\beta-\gamma} + o(|\delta| x^{-\beta-\gamma}), \tag{2.2}$$

uniformly in $x \geq 1$, where $\beta > 0$, $\gamma > 0$, $\delta = \delta(n)$ is nonrandom and $\delta \to 0$ as $n \to \infty$. [In the formula $\delta = \delta(n)$ we are using n as a surrogate for y_0; the value of n is of course a deceasing function of y_0.] Our aim is to predict the largest of

the next m values coming from distribution function (2.2), that is, to predict the largest of the next m values which exceed the given high threshold y_0 and come from distribution function (2.1).

It is common to take y_0 to be one of the data Y_i, in which case it would be included in $\mathcal{X}$ as the smallest value of X_i. On other occasions y_0 would be genuinely deterministic, for example, a threshold set on the machine that recorded the data. The value of y_0 would usually be taken large, relative to the center of the distribution of Y, so as to render small the value of δ at (2.2).

Often in practice the value of N is not known and is not itself of interest. For similar reasons, the distribution of Y is often not of interest, and typically there are no data in the lower tail. (The operation of conditioning on n sometimes amounts to defining the extrema of interest to be those that exceed the threshold y_0.) We shall assume that this is the case and regard n as a deterministic quantity. Our goal, then, is to determine the impact that the size of n has on coverage accuracy and on the extent of extrapolation that is feasible.

Of course, one could include the number, n, of data Y_i that exceed y_0, as well as the values of those data, in the procedure for inference. This is seldom done in practice, however, because little useful information is contained in n. In the exact Pareto case, where (2.1) would be replaced by $P(Y > y) = a_1 y^{-\beta}$, there is hardly more information in n than could be obtained by including one more data value—the largest datum strictly less than y_0. In the more general context of (2.1) the distribution of n, given $y_0, Y_1, \ldots, Y_n$, depends on β only through the high-order terms in (2.1), which as we shall see in Section 2.3 are very hard to obtain information about. Therefore, even in the general case it is conventional to condition on n.

2.2. *Prediction intervals for future events.* The basic Pareto model is the case $\delta = 0$ in (2.2): $P(X_i > x) = x^{-\beta}$, $x \geq 1$. There, if $\hat{\beta}_H$ denotes the Hill (1975) estimator of β, that is, $\hat{\beta}_H^{-1} = n^{-1} \sum_i \log X_i$, it may be shown that an exact prediction interval (i.e., an interval having exactly known coverage probability) for the largest of the next m values of X is $(1, \hat{x}_1(\alpha, m))$, where $\hat{x}_1(\alpha, m) = \rho^{-1/\hat{\beta}_H}$ and $\rho = \rho(\alpha, m)$ denotes the solution of the equation $\Psi_m(\rho) = \alpha$, and

$$\Psi_m(\rho) = \sum_{j=0}^{m} \binom{m}{j} (-1)^j (1 - n^{-1} j \log \rho)^{-n}. \tag{2.3}$$

Since Ψ_m is a strictly decreasing function of $\rho \in (0, 1)$, decreasing from 1 at $\rho = 0$ to 0 at $\rho = 1$, then $\rho(\alpha, m)$ is always uniquely defined. Outline derivations of these properties will be given in Section 4.1.

In many instances, for example, in the context of sampled distributions with regularly varying tails, it is attractive to use the Pareto prediction interval $(1, \hat{x}_1(\alpha, m))$ even when the Pareto model is not strictly correct. In such cases we are concerned to know the level of coverage error that is incurred. To address this

question we note that if F denotes the true distribution of X, then the probability that the largest of the next m values of X does not exceed $\hat{x}_1(\alpha, m)$ equals

$$\pi(\alpha, m) = E\big[F\{\hat{x}_1(\alpha, m)\}^m\big].$$

In the Pareto case, that is, when $P(X \geq x) = x^{-\beta}$, we have $\pi(\alpha, m) = \alpha$, but more generally [e.g., under the model (2.2)], interest centers on the extent of coverage error and the manner in which it alters as m increases. Our first theorem focuses on these issues.

THEOREM 2.1. *Assume* (2.2), *in which* $\beta > 0$, $\gamma > 0$ *and* $\delta \to 0$ *as* $n \to \infty$, *and that for some* $C > 0$, $n^{-C} = O(|\delta|)$.

(i) *If* m *is fixed*, *then*

$$\pi(\alpha, m) = \alpha + m\delta\rho(1-\rho)^{m-1}\left(\frac{\gamma \log \rho}{\beta + \gamma} + 1 - \rho^{\gamma/\beta}\right) + o(|\delta|) \tag{2.4}$$

as $n \to \infty$.

(ii) *If* $m = m(n) \to \infty$ *and* $(\delta + n^{-1/2}\log n)\log m \to 0$ *as* $n \to \infty$, *then*

$$\pi(\alpha, m) = \alpha + \frac{\alpha\delta\gamma \log \alpha}{\beta + \gamma} \log m + o(|\delta| \log m) \tag{2.5}$$

as $n \to \infty$.

REMARK 2.1 (Size of coverage error). It follows from Theorem 2.1 that the coverage error of the prediction interval $(1, \hat{x}_1(\alpha, m))$ is of size δ for fixed m, and that it increases only logarithmically fast, that is, at the rate $|\delta| \log m$, as $m \to \infty$.

Note too that m may increase exponentially fast, as a function of δ^{-1}, without these results being violated and without the prediction interval $(1, \hat{x}_1(\alpha, m))$ failing to have asymptotically correct coverage. For example, if $\delta = O(n^{-\xi_1})$ for some $\xi_1 \in (0, \frac{1}{2})$, then, in order for $\pi(\alpha, m)$ to converge to α as $n \to \infty$, it is sufficient that $m = O\{\exp(Cn^{\xi_2})\}$ for some $\xi_2 \in (0, \xi_1)$ and some $C > 0$.

REMARK 2.2 (Sign of coverage error). Observe that, since $\log \alpha < 0$, the sign of the term of size $\delta \log m$ on the right-hand side of (2.5) is opposite to the sign of δ. Moreover, for fixed m the sign of the dominant term (other than α) on the right-hand side of (2.4) is also opposite to the sign of δ. This follows from the fact that $(\beta + \gamma)^{-1}\gamma \log \rho + 1 - \rho^{\gamma/\beta} < 0$ for all $\gamma > 0$.

2.3. *Correcting coverage error.* In this section, we suggest three methods for improving coverage accuracy under a specific model assumption. The dominant terms in the coverage error formulas (2.4) and (2.5) equal δ multiplied by a factor that depends on the sampled distribution only through the ratio β/γ. Exploratory and diagnostic methods are available for assessing the value of this quantity

[Feuerverger and Hall (1999)]. Thus, after analysis of the data we may be prepared to assume a specific value for β/γ. The value $\beta/\gamma = 1$ is of particular interest in this regard; if X is generated in the form $X = |U|^{1/\beta}$ or $X = \mathrm{sgn}(U)|U|^{1/\beta}$, where the random variable U has a density that is nonzero and differentiable at the origin, then $\beta/\gamma = 1$. This value also obtains if X has a Type I extreme value distribution, that is, if $P(X \le x) = \exp(-cx^{-\beta})$, where $\beta, c > 0$.

In some contexts, β/γ may be explicitly estimated by least-squares or maximum likelihood, for example, by fitting the model consisting of the first two terms on the right-hand side of (2.2), that is,

$$P(X > x) = (1 - \delta)x^{-\beta} + \delta x^{-\beta-\gamma},$$

or by fitting a model that is asymptotically equivalent, to second order, to this one. These approaches were discussed by Feuerverger and Hall (1999). Experience in that setting suggests that estimation of β/γ is infeasible unless n is very large, and that $\mathcal{X}$ often contains little information about second-order aspects of an extreme-value model. In particular, for the sample sizes of 40–50 treated in the numerical work in the present paper, it is generally not feasible to estimate γ.

Statistical, mathematical and computational issues depend only a little on the fixed value chosen for β/γ. Since $\beta/\gamma = 1$ is more commonly encountered in practice (see the discussion two paragraphs above) we shall assume that case here. Then (2.4) reduces to

$$\pi(\alpha, m) = \alpha + m\delta\rho(1 - \rho)^{m-1}\big(\tfrac{1}{2}\log\rho + 1 - \rho\big) + o(|\delta|), \tag{2.6}$$

and (2.5) reduces to

$$\pi(\alpha, m) = \alpha + \tfrac{1}{2}\alpha\delta\log\alpha\log m + o(|\delta|\log m). \tag{2.7}$$

Our three methods for improving coverage accuracy are based on fitting a mixture of Pareto distributions,

$$G(x \mid \beta, \delta) = (1 - \delta)(1 - x^{-\beta}) + \delta(1 - x^{-2\beta}), \qquad x \ge 1, \tag{2.8}$$

to data, using either maximum likelihood or least squares. This aspect of the technique is a variant of that discussed by Feuerverger and Hall (1999). In (2.8), $\beta > 0$ and $-1 \le \delta < 1$, and we may construct maximum likelihood estimators $\hat\beta$ and $\hat\delta$, say. Note the distinction between $\hat\beta$, the first component of the maximum likelihood estimator under (2.8), and $\hat\beta_H$, the maximum likelihood estimator under the model $G(x|\beta) = 1 - x^{-\beta}$; we use the former in all three methods.

1. *First method*: *explicit error correction*—Let $\rho = \rho(\alpha, m)$ be defined as in Section 2.2, and put $p(\alpha, m) = m\rho(1 - \rho)^{m-1}(\frac{1}{2}\log\rho + 1 - \rho)$. Define $a = a_1 = a_1(\alpha, m, \hat\delta)$ to be the solution of $a + \hat\delta p(a, m) = \alpha$, and put $\hat x_2(\alpha, m) = \hat x_1\{a_1(\alpha, m, \hat\delta), m\}$. Then our error-corrected prediction interval is $(1, \hat x_2(\alpha, m))$.

2. *Second method*: *bootstrap calibration of Pareto-mixture interval*—Let $\hat{x}_1^*(a, m)$ denote the version of $\hat{x}_1(a, m)$ computed for a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ drawn from the model $G(\cdot|\hat{\beta}, \hat{\delta})$, rather than for the original data set $\mathcal{X} = \{X_1, \ldots, X_n\}$. Let $\hat{a}_2 = \hat{a}_2(\alpha, m, \mathcal{X})$ be the solution of $E[G^m\{\hat{x}_1^*(a, m)|\hat{\beta}, \hat{\delta}\}|\mathcal{X}] = \alpha$. Then, our error-corrected prediction interval is $(1, \hat{x}_1(\hat{a}_2, m))$.
3. *Third method*: *bootstrap calibration of naive interval*—Let $x_2 = x_2(\alpha, m, \beta, \delta)$ denote the solution of $G^m(x_2|\beta, \delta) = \alpha$. The "naive" or "empirical" α-level prediction interval for the next m values of X is $(1, x_2(\alpha, m, \hat{\beta}, \hat{\delta}))$ and may be shown to have coverage error $O\{n^{-1}(1 + \log m)\}$ when the model $G(\cdot \mid \beta, \alpha)$ is correct. This error may be reduced to $O\{n^{-2}(1 + \log m)^2\}$ by calibration, as follows. Let $(\hat{\beta}^*, \hat{\delta}^*)$ denote the version of $(\hat{\beta}, \hat{\delta})$ computed from $\mathcal{X}^*$ (defined as for the second method) rather than from $\mathcal{X}$, and let $\hat{a}_3 = \hat{a}_3(\alpha, m, \mathcal{X})$ be the solution of

$$E\big[G^m\{x_2(a, m, \hat{\beta}^*, \hat{\delta}^*) \mid \hat{\beta}, \hat{\delta}\}|\mathcal{X}\big] = \alpha.$$

Then our error-corrected prediction interval is $(1, x_2(\hat{a}_3, m, \hat{\beta}, \hat{\delta}))$.

The first method, which does not involve any simulation and so is in principle the simplest of the three, is strongly asymptotic in character and can share the difficulties associated with related techniques for coverage correction. In particular, the approximations

$$\alpha + m\delta\rho(1 - \rho)^{m-1}\big(\tfrac{1}{2}\log\rho + 1 - \rho\big) \quad \text{and} \quad \alpha + \tfrac{1}{2}\alpha\delta\log\alpha\log m$$

to coverage probability, deriving from (2.6) and (2.7) respectively, do not necessarily lie in the interval $[0, 1]$. This can lead to problems when solving the equation that defines $a_1(\alpha, m, \delta)$. (Similar difficulties arise in the context of explicit Edgeworth correction.) For these reasons we prefer the second and third methods.

However, it should be recognized that all three approaches introduce a new term to the coverage error formula, of size $O\{n^{-1}(\log m)^2\}$ for methods 1 and 3 and $O\{n^{-2}(\log m)^4\}$ in the case of method 2. These errors occur because a stochastic error of order $\eta = n^{-1/2}\log m$ arises from estimating δ in the model (2.8), even when the model is correct and $\delta = 0$. Since the expected value of the error is of order η^2 this is the order of the additional coverage-error term for methods 1 and 3. It is reduced to $O(\eta^4)$ by the calibration involved in method 2. As a result, none of the methods 1, 2 or 3 exhibits perfect coverage accuracy when the sampled distribution is exactly Pareto.

We conclude with a formal description of these results.

THEOREM 2.2. *Assume* (2.8) *holds for* $\beta = \gamma$. *Then the coverage probabilities of prediction intervals produced by the first and third methods equal* $\alpha + O\{n^{-1}(1 + \log m)^2\} + o\{|\delta|(1 + \log m)\}$, *while for the second method they equal* $\alpha + O\{n^{-2}(1 + \log m)^4\} + o\{|\delta|(1 + \log m)\}$.

In the event that the assumption $\beta/\gamma = 1$ is incorrect, the assertions of the theorem do not hold. Instead, taking the setting of Theorem 2.1(iii) as an example, and using the first or the third method to "correct" for coverage error, we have instead of the result given in Theorem 2.2 the property that the coverage probability equals

$$\alpha + \frac{\alpha\delta(\gamma - \beta)\log\alpha}{2(\beta + \gamma)} \log m + O\{n^{-1}(\log m)^2\} + o(\delta \log m).$$

Therefore, there is no improvement, and possibly even a deterioration, in the order of coverage error asserted by Theorem 2.1(ii).

We are not claiming that any of our methods for improving coverage accuracy is "optimal" in some sense. Indeed, the semiparametric nature of the setting where the methods are deployed and the fact that the coverage corrections address second-order quantities rather than first-order ones mean that it is probably not feasible to describe optimality in a manner that is meaningful for practical applications.

The technique of proof of Theorem 2.2 is a combination of those used to derive Theorems 2.1 and 3.2, and so is not given here.

2.4. *Simulation study.* In this section we investigate the coverage accuracies and lengths of both the Pareto interval and its bootstrap-calibrated version, the latter constructed using the second method (referred to below as Method 2) suggested in Section 2.3. We took the true distribution to be given by model (2.8). For reasons given in Section 2.3, intervals constructed using Method 2 are generally more satisfactory than those based on either of the other two calibrated techniques.

Coverage accuracy was estimated by averaging over 200 realizations of sample size $n = 50$ from the population with distribution function given by (2.8). We took $\beta = 1$ and either $\delta = n^{-1/5}$ or $\delta = n^{-2/5}$. In the bootstrap calibration step we used 400 simulations for each of the 200 samples.

In Figure 1 the "true" coverage of 90% prediction intervals is plotted against m. Figure 2 shows the values of $\hat{x}_1(\alpha, m)$ and $\hat{x}_1(\hat{a}_2, m)$, as functions of m, averaged over all 200 samples; these are of course numerical approximations to the expected values of $\hat{x}_1(\alpha, m)$ and $\hat{x}_1(\hat{a}_2, m)$, respectively. The latter figure also shows the solution x_0 of the equation $G(x_0)^m = \alpha$.

From Figure 1 we see that, except when m is small, coverage accuracy of both interval types tends to decrease in a logarithmic way as m increases; and that the rate of decrease is faster for the Pareto interval $(1, \hat{x}_1(\alpha, m))$ than for its Method 2 counterpart. Figure 2 shows that this is achieved through an increase in the average length of the prediction region, as a function of m. The reason the basic Pareto interval fails to achieve good coverage accuracy, for large m, is that its length does not increase sufficiently fast with m.

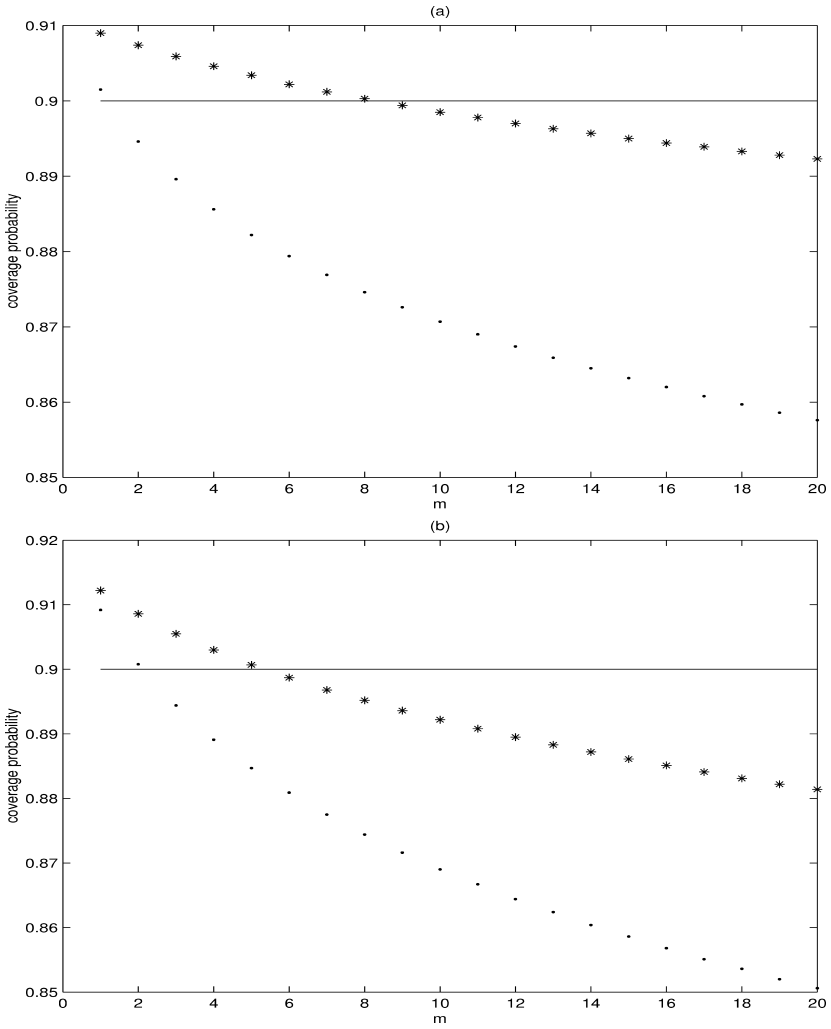

FIG. 1. *Coverage accuracy*: *true coverages of one-sided prediction intervals* $(1, \hat{x}_1(\alpha, m))$ *and* $(1, \hat{x}_1(\hat{a}_2, m))$, *for* $\alpha = 0.90$, *are indicated by dot and star points*, *respectively*. *The unbroken line indicates the true coverage level*. *Panels* (a) *and* (b) *show coverages when* $\delta = n^{-2/5}$ *and* $\delta = n^{-1/5}$, *respectively*.

As expected, the extent to which coverage accuracy decreases with increasing m is greater for the larger value of δ, although Figure 1 indicates that the differences are only small. Figure 2 confirms that the greater extent of undercoverage observed for larger δ is caused by the relatively large gap between the too-short length of the corresponding prediction interval and the length of its "ideal," nonempirical counterpart, based on x_0. However, this is only in absolute terms; in relative terms the lengths of the empirical prediction intervals in the cases $\delta = n^{-1/5}$ and $\delta = n^{-2/5}$ are too small by similar amounts.

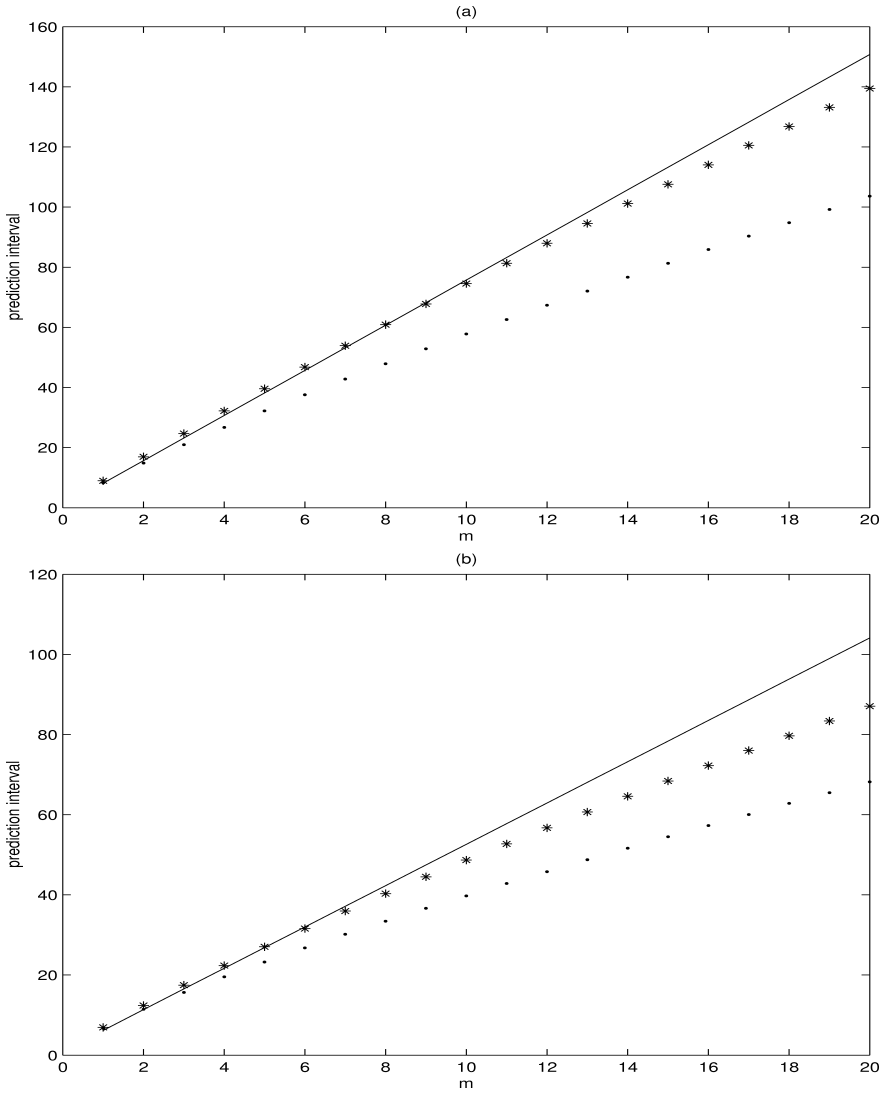

FIG. 2. *Average sizes of upper endpoints*: *values of* $E\{\hat{x}_1(\alpha, m)\}$ *and* $E\{\hat{x}_1(\hat{a}_2, m)\}$, *for* $\alpha = 0.90$, *are indicated by dot and star points*, *respectively*. *The straight line shows the solution*, x_0, *of the equation* $G(x_0)^m = \alpha$. *Panels* (a) *and* (b) *show expected values when* $\delta = n^{-2/5}$ *and* $\delta = n^{-1/5}$, *respectively*.

3. Parametric prediction in a generalized Pareto model.

3.1. *Statement of problem.* In Section 3 we assume that the form of departure from the Pareto distribution is known exactly, up to an unknown parameter. We complement the results in Section 2 by discussing the effects that long-range forecasting has on accuracy in this setting, when bootstrap methods are used for

calibration. The conclusions are broadly similar (i.e., extrapolation can be made exponentially far into the future before serious errors are introduced) but the quantifications of the errors are quite different in the two settings. The generalized Pareto model treated here provides an illustration of the accuracy of bootstrap-calibrated forecasts under exact models, just as the Pareto-type approximations in Section 2 illustrate forecasting accuracy under distributional approximations. In Remark 3.1 we address bootstrap calibration of a misspecified generalized Pareto model.

Suppose we observe events with strengths $X_1, \ldots, X_n$ at respective times $T_1 \le \cdots \le T_n$. Here, we have the following properties: (i) the variables in $\mathcal{X} = \{X_1, \ldots, X_n\}$ are independent and have common distribution $F(x|\theta) = 1 - (1 + \tau x)^{-\beta}$, for $x \ge 0$, where $\beta > 0$ and $\tau > 0$ are parameters, and $\theta = (\beta, \tau)^T$; (ii) the variables in $\mathcal{T} = \{T_1, T_2 - T_1, \ldots, T_n - T_{n-1}\}$ are independent and exponentially distributed, with common mean $\mu > 0$; and (iii) the variables in $\mathcal{X}$ are independent of those in $\mathcal{T}$.

Properties (i)–(iii) define events observed at successive points of a Poisson process in time. Assumption (i) that $\beta > 0$, which we make throughout our work, ensures that we are addressing the so-called regular case where maximum likelihood estimators of β and τ are jointly asymptotically Normally distributed; see Smith (1987).

The distribution $F(\cdot|\theta)$ is usually referred to as the generalized Pareto distribution (GPD); see, for example, Embrechts, Klüppelberg and Mikosch [(1997), pages 162ff.]. Of course, we may allow τ to vary with n without influencing our results. Changing τ amounts only to altering scale, and it is the value of $(\hat{\tau} - \tau)/\tau$, rather than simply $\hat{\tau} - \tau$, which determines forecasting accuracy. Embrechts, Klüppelberg and Mikosch [(1997), page 165] and Reiss and Thomas [(1997), page 54] have discussed the fact that varying τ with n in the GPD model can be interpreted as modelling exceedences over high thresholds.

We shall consider two distinct prediction problems, where we wish to construct an α-level prediction interval for (a) the largest of the strengths, $X_{\max}(m)$, of the next m events or (b) the strength, $X_{\max}[t]$, of the strongest further event that occurs in the next time period of length t units. One-sided intervals for solving these problems, and having nominal levels α, are respectively

$$\begin{aligned} &\text{(a)} \quad \mathcal{I}(\alpha, m) = \big[0, \widetilde{F}^{-1}(1 - \alpha^{1/m})\big], \\ &\text{(b)} \quad \mathcal{I}[\alpha, t] = \big[0, \widetilde{F}^{-1}\{(-\log \alpha)\widehat{\mu}/t\}\big], \end{aligned} \tag{3.1}$$

where $\widehat{\mu} = T_n/n$, $\widetilde{F} = 1 - F(\cdot|\hat{\theta})$ and $\hat{\theta}$ denotes the maximum likelihood estimator of θ.

These are the so-called naive or estimative prediction intervals for problems (a) and (b), obtained by substituting parameter estimates for true parameter values in the intervals that would be employed if true parameter values were known.

For fixed m or t, and without calibration, the coverage accuracy of such a procedure is only $O(n^{-1})$. The order of accuracy is not improved, however, using the methods of predictive likelihood; for the latter, see, for example, Butler (1986), Davison (1986), Bjørnstad (1990) and Barndorff-Nielsen and Cox [(1994), Section 9.4], and note the discussion by Hall, Peng and Tajvidi (1999) of coverage accuracy of predictive likelihood. The GPD model can, however, be addressed using predictive pivotal methods suggested by Barnard (1986) in location–scale cases.

3.2. *Effects of long-range prediction on coverage accuracy.* Of particular interest is the effect on coverage accuracy of increasing the values of m and t. One may show that in order for coverage to converge to the nominal level, α, as $n \to \infty$, it is necessary and sufficient that neither m nor t increases faster than $\exp\{o(n^{1/2})\}$, as our next result indicates.

THEOREM 3.1. *If m and t are bounded above* 0, *then in the case of problems* (a) *and* (b) *a necessary and sufficient condition for the prediction intervals $\mathcal{I}(\alpha, m)$ and $\mathcal{I}[\alpha, t]$ to have asymptotically correct coverage, for each $0 < \alpha < 1$, is that* $\log m + \log t = o(n^{1/2})$ *as* $n \to \infty$.

It may be proved that this result also holds for bootstrap-calibrated forms of the intervals $\mathcal{I}(\alpha, m)$ and $\mathcal{I}[\alpha, t]$, which we consider next. Let $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ denote a resample drawn by sampling randomly from the distribution $F(\cdot|\hat{\theta})$, conditional on $\mathcal{X}$; and conditionally independently of $\mathcal{X}^*$, let $\mathcal{T}^* = \{T_1^*, T_2^* - T_1^*, \ldots, T_n^* - T_{n-1}^*\}$ be a set of independent Exponential random variables with mean $\widehat{\mu}$. Let $\mathcal{I}^*(\alpha, m)$ and $\mathcal{I}^*[\alpha, t]$ [denoted generically by $\mathcal{I}^*(\alpha)$] be the bootstrap versions of $\mathcal{I}(\alpha, m)$ and $\mathcal{I}[\alpha, t]$ [denoted generically by $\mathcal{I}(\alpha)$], respectively, obtained by replacing $(\mathcal{X}, \mathcal{T})$ by $(\mathcal{X}^*, \mathcal{T}^*)$ in definitions of the latter prediction intervals.

In the context of $\mathcal{I}^*(\alpha, m)$ or $\mathcal{I}^*[\alpha, t]$, respectively, write $X_{\max}^*$ for the largest of the next m bootstrap values of strength, or of the bootstrap values of strength observed in the next time interval of length t. Put $\hat{\pi}(\alpha) = P\{X_{\max}^* \in \mathcal{I}^*(\alpha)|\mathcal{X}\}$, the bootstrap estimator of $\pi(\alpha) = P\{X_{\max} \in \mathcal{I}(\alpha)\}$. Define $\gamma = \widehat{\gamma}_\alpha$ to be the solution of $\hat{\pi}(\gamma) = \alpha$. Then the bootstrap-calibrated form of $\mathcal{I}(\alpha)$ is $\mathcal{I}(\widehat{\gamma}_\alpha)$, denoting either $\mathcal{I}(\widehat{\gamma}_\alpha, m)$ or $\mathcal{I}[\widehat{\gamma}_\alpha, t]$.

We may iterate this approach any number, k, of times, obtaining intervals $\mathcal{I}_k(\alpha, m)$ and $\mathcal{I}_k[\alpha, t]$, say. In particular, $\mathcal{I}_1(\alpha, m) = \mathcal{I}(\widehat{\gamma}_\alpha, m)$ and $\mathcal{I}_1[\alpha, t] = \mathcal{I}[\widehat{\gamma}_\alpha, t]$. We claim that each iteration reduces coverage error by an order of magnitude and that coverage error increases only logarithmically fast as a function of the distance into the future at which we are making our prediction. Our next result makes this clear.

THEOREM 3.2. *Assume that, as* $n \to \infty$, *both* $t = t(n)$ *and* $m = m(n)$ *are bounded away from* 0 *and satisfy* $\log t + \log m = O(n^{(1/2)-\varepsilon})$, *for some* $0 < \varepsilon < \frac{1}{2}$. *Then*

$$\begin{aligned} P\{X_{\max}(m) \in \mathcal{I}(\alpha, m)\} &= \alpha + O\big[\{n^{-1}(1+\log m)^2\}^k\big], \\ P\{X_{\max}[t] \in \mathcal{I}[\alpha, t]\} &= \alpha + O\big[\{n^{-1}(1+|\log t|)^2\}^k\big]. \end{aligned} \tag{3.2}$$

REMARK 3.1 (Comparison with the Pareto case in Section 2). There are similarities between (3.2) and the coverage expansions (2.4) and (2.5) in the Pareto case. For example, in both the effect of increasing m is only to inflate coverage error by a power of $\log m$. However, in the semiparametric context of Section 2 the power is 1, whereas in the parametric setting of Theorem 3.2 it is an even integer, depending on iteration order. The reason for the difference is that in Section 2 the coverage error resulted largely from model misspecification—we assumed a Pareto distribution when the correct model was only asymptotically Pareto—while in Theorem 3.2 the error comes from calibrating a "naive" or "estimative" interval in a strictly parametric setting.

Related arguments may be used to address the "interaction" between contexts treated in Sections 2 and 3, that is, where the GPD model is mistakenly assumed, when it is not valid, and predictions are made using bootstrap methods rather than the technique (derived from a standard Pareto, instead of a GPD, model) discussed in Section 2. It may be shown that, to first order, the results in this setting are just the compound, through addition, of those in Sections 2 and 3.

That is, the order of magnitude of the coverage error of a prediction interval for $X_{\max}(m)$ is the order stated in Theorem 2.1, plus that in result (3.2) of Theorem 3.2; the former is $O(\delta \log m)$ and the latter equals $O[\{n^{-1}(1+\log m)^2\}^k]$. In this setting the data X_i are assumed to satisfy, instead of (2.2), the condition

$$P(X_i > x) = (1-\delta)(1+\tau x)^{-\beta} + \delta(1+\tau_1 x)^{-\beta-\gamma} + o\big\{|\delta|(1+x)^{-\beta-\gamma}\big\}$$

uniformly $x > 0$, where $\tau_1 > 0$ and $\delta = \delta(n) \to 0$; and the bootstrap method, based on the GPD assumption, is used to construct prediction intervals.

REMARK 3.2 (Analogous results for confidence intervals). Standard bootstrap methods, for example, the percentile method, may be used to construct confidence intervals involving long-range extrapolation. In particular, we might seek a confidence interval for (i) the mean interoccurrence time $\nu(x|\mu,\theta) = \mu/\bar{F}(x|\theta)$ (where $\bar{F} = 1 - F$) of successive events of strength at least x; or for (ii) the p-level quantile of the distribution of strength; or for (iii) the probability that strength exceeds a given value, x.

Bootstrap calibration may then be employed to improve coverage accuracy, and results analogous to those in Theorem 3.2 may be derived in this setting. For example, the k-fold bootstrap iterate of a one-sided percentile-method confidence interval may be shown to have coverage error $O\{(n^{-1/2}\log u)^k\}$, where $u = x$,

p^{-1} and x in the cases of problems (i), (ii) and (iii), respectively. In the setting of two-sided, k-fold bootstrap-calibrated, percentile-method confidence intervals the coverage error is $O[\{n^{-1}(\log u)^2\}^k]$.

The main feature of these results, common to Theorems 2.1 and 3.2, is the way in which coverage error increases only logarithmically with extrapolation. The property that the size of coverage error decreases with k is common in this type of problem; see Hall [(1992), Chapter 3] for discussion. However, the constant multiple of the asymptotic order can also increase with k, and as a rule, for a given sample size, there is a finite value of k for which the maximum amount of correction is achieved.

REMARK 3.3 (Two-sided prediction intervals). In extreme-value applications, one-sided intervals such as those addressed in Theorem 3.2 are usually of more interest than their two-sided counterparts. Nevertheless, the theorem holds without change for equal-tailed, two-sided prediction intervals formed by taking the intersection of two one-sided intervals. In both one- and two-sided cases the "big oh" remainder terms correctly reflect the size of the remainder, although, particularly in the case where m and t are bounded, exact formulae for dominant terms in remainders are complex.

The fact that there is no reduction in the order of coverage error in the two-sided case is in contrast to the context of confidence intervals, where parity properties of terms in Edgeworth expansions imply such a reduction. In the case of prediction intervals, however, expansions of coverage error are derived directly from Taylor expansions of the sampled distribution function, rather than by Edgeworth expansion.

3.3. *Numerical properties.* We present both an application to real data and a simulation study. The former is based on financial loss data supplied by the Swedish insurance group Länförsäkringar for the most severe windstorms during the 12-year period from 1 January 1982 to 10 July 1993. The "strength" of a storm is defined as the total insurance loss that it produced, in millions of Swedish Kronor (MSEK). A windstorm is defined as an "event" if its strength exceeds 0.9 MSEK. The time of occurrence of each event was deduced from wind speed data provided by the Swedish Meteorological and Hydrological Institute. Figure 3a graphs storm strength against time of occurrence.

These strength data, in a sample of size $n = 45$, fit a generalized Pareto distribution well; see the Q–Q plot in Figure 3b, and note the goodness-of-fit analysis by Rootzén and Tajvidi (1997), where the data set is discussed in detail. Parameter estimates are $\hat{\beta} = 1.354$, $\hat{\tau} = 0.199$ and $\widehat{\mu} = 95.4$.

We applied the methodology suggested in Sections 3.1 and 3.2 to these data, and illustrate here the solution, given in (3.1), to problem (b). Table 1 presents prediction intervals for $X_{\max}[t]$. Here and in the simulation study, when using the bootstrap to calibrate prediction intervals we employed 500 simulations. Overall,

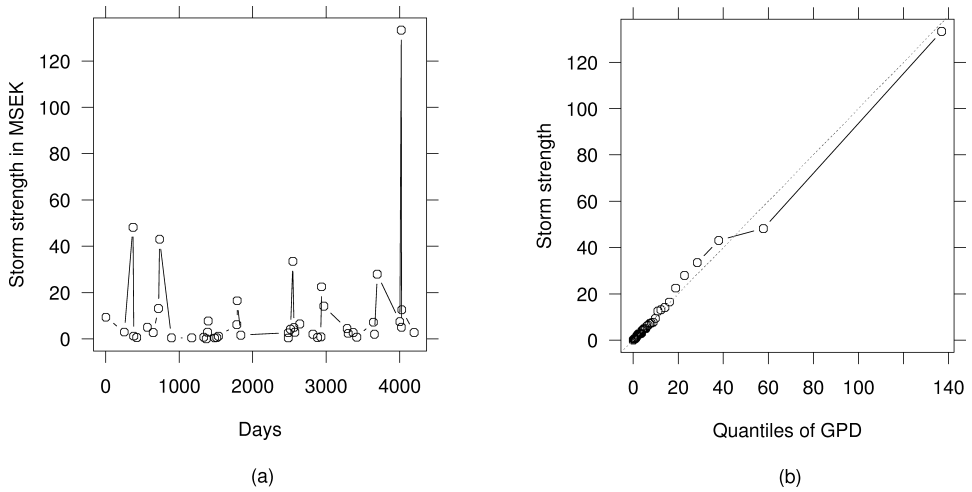

FIG. 3. *Windstorm data*: (a) *occurrence times of storms for which the strength exceeded MSEK* 0.9, *plotted on the horizontal axis*, *with the logarithm of strength* (*in MSEK*) *on the vertical axis*; (b) *a* Q–Q *plot for these data*, *the dashed line corresponding to the equation* $y = x$.

the interval endpoints in Table 1 increase by a factor of about 2.5 for each doubling of t, the latter measured in days.

Table 2 reports results from a simulation study that attempts to capture performance in the context noted immediately above. There we employed parameter values close to those observed for the real data; we used $n = 45$, $\beta = 1$, $\tau = 1$ and $\mu = 100$. (The results are of course invariant under changes to τ and μ.) Table 2 confirms that, even after significant extrapolation, bootstrap-calibrated prediction intervals enjoy particularly good coverage accuracy.

In these results and others not reported here we found that a single calibration step was adequate to ensure excellent robustness against extrapolation. Even

TABLE 1

Endpoints $\hat{x}_j$ *of prediction intervals* $\mathcal{I}_1$ (*one-sided lower-tailed*), $\mathcal{I}_2$ (*one-sided upper-tailed*) *and* $\mathcal{I}_3$ (*two-sided*) *for* $X_{\max}[t]$: *values of* t *are in the range* $t = 365, 1460(1460)7300$, *and nominal level is* 0.90. *For one-sided intervals* $\mathcal{I}_1$ *and* $\mathcal{I}_2$, *only upper or lower limits are listed*; *for two-sided intervals* $\mathcal{I}_3$, *both limits are given*

t	$\hat{x}_2$ for $\mathcal{I}_1$	$\hat{x}_1$ for $\mathcal{I}_2$	$\hat{x}_1$ for $\mathcal{I}_3$	$\hat{x}_2$ for $\mathcal{I}_3$
365	8.1×10^7	1.8×10^6	7.6×10^5	1.8×10^8
1460	3.5×10^8	1.5×10^7	1.0×10^7	1.3×10^9
2920	8.5×10^8	2.8×10^7	2.3×10^7	2.2×10^9
4380	1.3×10^9	3.7×10^7	2.9×10^7	2.8×10^9
5840	2.0×10^9	4.7×10^7	3.5×10^7	4.3×10^9
7300	2.5×10^9	5.5×10^7	3.9×10^7	5.3×10^9

TABLE 2
Monte Carlo approximations to true coverage of prediction intervals for $X_{\max}[t]$ in the case of a fitted GPD model: nominal levels are 0.90 *and* 0.95, *with* $n = 45$ *and* $\beta = 1$. *The value of* t *is given in the first column*

	$\alpha = 0.90$		$\alpha = 0.95$	
t	**no calib.**	**calib.**	**no calib.**	**calib.**
1500	0.838	0.905	0.897	0.954
3000	0.817	0.910	0.879	0.953
5000	0.798	0.916	0.863	0.950

greater accuracy is obtainable using the double bootstrap, but practical motivation for such a high order of correction does not seem strong, in view of the good performance evident from Table 2.

4. Technical arguments.

4.1. *Derivation of formula for exact prediction intervals in Pareto case.* Let $Y_1, \ldots, Y_m$ be independent and identically distributed with distribution function $F_\beta(x) = 1 - x^{-\beta}$, for $x > 1$; independently of the X_i's, let $Z_1, \ldots, Z_n$ be independent standard exponential variables; and recall the definition of $\Psi_m(\rho)$ in (2.3). Put $S = n^{-1} \sum_i Z_i$ and, without loss of generality, write $\hat\beta_H = \beta/S$. Then, using the representation of independent Pareto random variables as functions of independent standard exponential random variables $Y_1, \ldots, Y_{m+n}$, one may prove that

$$P\Bigl(\max_{1\le i\le m} X_i \le \rho^{-1/\hat\beta_H}\Bigr) = P\Bigl(\max_{1\le i\le m} Y_{n+i} \le -n^{-1}\log\rho \sum_{i=1}^n Y_i\Bigr)$$
$$= \frac{1}{\Gamma(n)} \int_0^\infty \{1 - \exp(n^{-1} w \log\rho)\}^m w^{n-1} e^{-w}\, dw.$$

Expanding $\{1 - \exp(n^{-1} w \log\rho)\}^m$ as a binomial, and carrying out the integration term by term, we deduce the exactness of the prediction interval suggested in Section 2.1. It may also be proved that $E(1 - \rho^{\beta/\hat\beta_H})^m = \Psi_m(\rho)$, from which it can be seen that Ψ_m is strictly decreasing.

4.2. *Proof of Theorem* 2.1. Let nV_n denote a random variable having the Gamma distribution with shape parameter n and unit scale. The distribution of $\beta/\hat\beta_H$, which we shall denote by W_n, and the distribution of $V_n - \delta\gamma(\beta+\gamma)^{-1}$ have almost identical large-sample properties, and our method will be based on this feature.

In the above notation, ρ is the solution of $E(1-\rho^{V_n})^m = \alpha$. If $W_n = -n^{-1}\sum_{1\le i\le n}\beta\log X_i$, where X_i has distribution function F with the property (2.2), then W_n has the moment generating function m_{W_n} given by (on changing the argument to ns for simpler notation),

$$m_{W_n}(ns) = \left(\frac{1-\delta}{1-s} + \frac{\delta\{1+o(1)\}}{1-s\beta(\beta+\gamma)^{-1}}\right)^n$$
$$= \frac{1}{(1-s)^n}\left(1 - \frac{\{1+o(1)\}\delta\gamma s/(\beta+\gamma)}{1-s\beta(\beta+\gamma)^{-1}}\right)^n.$$

It follows that, as $n\to\infty$, (a) $E(W_n) = 1 - \delta\gamma(\beta+\gamma)^{-1} + o(|\delta|)$, (b) the distribution of $U_n \equiv n^{1/2}(W_n - EW_n)$ converges to that of a standard normal variable and (c) $\sigma_n^2 \equiv \operatorname{var} U_n = 1 - 2\delta\beta\gamma(\beta+\gamma)^{-2} + o(|\delta|)$.

Assume for the time being that the "small oh" term in (2.2) may be dropped so that the distribution of X_i is given more simply by

$$1 - F(x) = (1-\delta)x^{-\beta} + \delta x^{-\beta-\gamma}, \tag{4.1}$$

where $\delta\to 0$ as $n\to\infty$. Then F has a density, and we can compare the density of U_n with that of $n^{1/2}(V_n - 1)$. Moreover, the moment generating function of U_n is given by

$$m_{U_n}(s) = (1 - sn^{-1/2})^{-n}\left(1 - \frac{\delta\gamma s/(\beta+\gamma)n^{1/2}}{1-s\beta\{(\beta+\gamma)n^{1/2}\}^{-1}}\right)^n \exp(-sEW_n n^{1/2})$$
$$= (1-sn^{-1/2})^{-n}\exp\left\{-sn^{1/2} + \frac{1}{2}(\sigma_n^2 - 1)s^2 + o(|\delta|)\right\},$$

where the second identity above holds uniformly in values of s for which $|s| \le \zeta(n)n^{1/6}$ for any given sequence $\zeta(n)\downarrow 0$. (We allow s to take complex values.)

Comparing this formula with that for the moment generating function $m_{\widetilde{V}_n}$ of $\widetilde{V}_n \equiv n^{1/2}(V_n - 1)$, we obtain

$$m_{U_n}(s) = m_{\widetilde{V}_n}(s)\exp\left\{\tfrac{1}{2}(\sigma_n^2 - 1)s^2 + o(|\delta|)\right\}, \tag{4.2}$$

again uniformly in $|s| \le \zeta(n)n^{1/6}$. Noting (4.2), and inverting the moment generating functions of U_n and $\widetilde{V}_n$ to obtain the respective densities, which we denote by f_{U_n} and $f_{\widetilde{V}_n}$, we deduce that

$$f_{U_n}(x) = \sigma_n^{-1} f_{\widetilde{V}_n}(x/\sigma_n)\{1+o(|\delta|)\}, \tag{4.3}$$

where the $o(|\delta|)$ remainder term is of that order uniformly in x satisfying $|x| \le n^{1/6}\zeta(n)$, for any given sequence $\zeta(n)\downarrow 0$. The inversion may be accomplished via the representation

$$f_{U_n}(x) = \frac{1}{2\pi i}\int_{(x/\sigma_n^2)-i\infty}^{(x/\sigma_n^2)+i\infty} e^{-xs} m_{U_n}(s)\,ds,$$

where $i = \sqrt{-1}$.

Taking $\ell_n = \log n$, and using (4.3), we obtain, provided the distribution of X_i is given by (4.1), the results

$$\begin{aligned}
&E(1-\rho^{W_n})^m \\
&\quad = E\{(1-\rho^{W_n})^m I(|U_n| \le \ell_n)\} + O\bigl(\exp[-\tfrac{1}{2}\ell_n^2\{1+o(1)\}]\bigr) \\
&\quad = E\bigl\{\bigl(1-\exp[(\log\rho)\{1+U_n n^{-1/2}+o(|\delta|/n^{1/2})\}]\bigr)^m I(|U_n| \le \ell_n)\bigr\} + o(|\delta|) \\
&\quad = \int_{-\ell_n}^{\ell_n} [1-\exp\{(\log\rho)(1+un^{-1/2})\}]^m \sigma_n^{-1} f_{\widetilde{V}_n}(u/\sigma_n)\,du + o(|\delta|) \\
&\quad = \int_{-\infty}^{\infty} [1-\exp\{(\log\rho)(1+un^{-1/2})\}]^m f_{\widetilde{V}_n}(u)\,du + o(|\delta|) \\
&\quad = E\{(1-\rho^{V_n})^m\} + o(|\delta|) = \alpha + o(|\delta|).
\end{aligned}$$

The result of Theorem 2.1 follows, provided the "small oh" term in (2.2) may be taken to be identically 0. However, once we have the result in that case we also have it for the more general F of (2.2), as may be seen using a comparison argument based on stochastic ordering of distributions.

4.3. *Proof of Theorem* 3.1. Put $\bar{F} = 1 - F(\cdot|\beta,\tau)$, where β, τ denote the true values of those parameters, and let $\hat{\theta} = (\hat{\beta}, \hat{\tau})$, $\Delta_\beta = \hat{\beta} - \beta$, $\Delta_\tau = \hat{\tau} - \tau$, $\Delta_\mu = (\widehat{\mu} - \mu)/\mu$, $z = 1 + \tau x$ and $y = x/z$. Then

$$\widetilde{F}(x)/\bar{F}(x) = (1+y\Delta_\tau)^{-\beta}(z + x\Delta_\tau)^{-\Delta_\beta}, \tag{4.4}$$

where $\widetilde{F} = 1 - F(\cdot|\hat{\beta}, \hat{\tau})$. From (4.4) and the asymptotic normality of $\hat{\theta}$ [see Smith (1987)], if $x = x(n) \to \infty$ as $n \to \infty$, then $\widetilde{F}(x)/\bar{F}(x) \to 1$ in probability as $x \to \infty$ if and only if $\log x = o(n^{1/2})$. Likewise, if $m = m(n)$ is increasing, then $\widetilde{F}^{-1}(1-\alpha^{1/m})/\bar{F}^{-1}(1-\alpha^{1/m}) \to 1$ in probability as $m \to \infty$ if and only if $\log m = o(n^{1/2})$, and likewise it may be proved that if $t = t(n)$ is increasing, then

$$\widetilde{F}^{-1}\{(-\log\alpha)\widehat{\mu}/t\}/\bar{F}^{-1}\{(-\log\alpha)\mu/t\} \to 1$$

in probability as $t \to \infty$ if and only if $\log t = o(n^{1/2})$. Using these properties in conjunction with the definitions of prediction intervals we may deduce the claimed results.

For example, consider the prediction interval $\mathcal{I}(\alpha, m)$, where $m = m(n) \to \infty$, and put $u = u(n) = 1 - \alpha^{1/m}$ and $V = \widetilde{F}^{-1}(u)$ (a random variable). Note that the right-hand side of (4.4) equals $\{1 + o_p(1)\}(\tau x)^{-\Delta_\beta}$ as $x \to \infty$ and that $P(V > v) \to 1$ as $n \to \infty$, for each fixed $v > 0$. Therefore, by (4.4),

$$u = \widetilde{F}(V) = \{1+o_p(1)\}\bar{F}(V)(\tau V)^{-\Delta_\beta} = \{1+o_p(1)\}\tau^{-(\beta+\Delta_\beta)} V^{-(\beta+\Delta_\beta)}.$$

It follows that, on defining $\Delta \equiv \Delta_\beta/\beta$, we have $\tau V u^{1/(\beta+\Delta_\beta)} \to 1$ in probability and

$$\begin{aligned}
\bar{F}(V) &= \{1+o_p(1)\}u^{1/(1+\Delta)} \\
&= -\{1+o_p(1)\}(\log\alpha)\exp\{-(1+\Delta)^{-1}\log m\}.
\end{aligned} \tag{4.5}$$

In view of (4.5), the coverage probability of the prediction interval $\mathcal{I}(\alpha, m)$ equals

$$\begin{aligned}\pi(\alpha) &\equiv P\{X_{\max}(m) \le \widetilde{F}^{-1}(1-\alpha^{1/m})\} = E[\{1-\bar{F}(V)\}^m] \\ &= E\big([1+m^{-1}\{1+o_p(1)\}(\log\alpha)\exp\{\Delta(1+\Delta)^{-1}\log m\}]^m\big) \\ &= E\big(\exp[\{1+o_p(1)\}(\log\alpha)\exp\{\Delta(1+\Delta)^{-1}\log m\}]\big) \\ &= E\big\{\exp\big(\{1+o_p(1)\}(\log\alpha)\exp[\{1+o_p(1)\}N\eta]\big)\big\},\end{aligned} \tag{4.6}$$

where $N \equiv n^{1/2}\Delta$ is asymptotically Normal $N(0,\sigma^2)$ for some $\sigma > 0$, and $\eta \equiv n^{-1/2}\log m$.

Since $1 - \bar{F}(V) \le 1$ the argument of the expectation is uniformly bounded by 1. It follows from this property and (4.6), on using a subsequence argument (i.e., considering the case where, along a subsequence of values n, $\eta = \eta(n) \to \ell$, with $0 \le \ell \le \infty$), that $\pi(\alpha) \to \alpha$ for each $0 < \alpha < 1$ if and only if $\eta \to 0$. That is, the prediction interval $\mathcal{I}(\alpha, m)$ has asymptotically correct coverage, for all α, if and only if $n^{-1/2}\log m \to 0$. [If $\eta \to \infty$, then $\pi(\alpha) \to \frac{1}{2}$ for each α, and if $\eta \to \ell \in (0,\infty)$, then $\pi(\alpha) \to \pi_0(\alpha) \equiv E[\exp\{(\log\alpha)e^{N_0\ell}\}]$, where N_0 is Normal $N(0,\sigma^2)$. The value of $\pi_0(\alpha)$, depending as it does on α and $\ell\sigma$, can equal α or be on either side of α. Therefore, the condition $n^{-1/2}\log m \to 0$ is not equivalent to $\pi(\alpha) \to \alpha$ for a single, given value of α.]

4.4. *Proof of Theorem* 3.2. Let $u = 1 - \alpha^{1/m}$, $V = V(\alpha) = \widetilde{F}^{-1}(u)$, $v = v(\alpha) = \bar{F}^{-1}(u)$, $Y = V/(1+\tau V)$ and $W = \log(1+\tau V)$. In view of (4.4), V is equivalently defined by

$$u(1+Y\Delta_\tau)^{\beta+\Delta_\beta}\exp(W\Delta_\beta)(1+\tau V)^\beta = 1. \tag{4.7}$$

Use Taylor expansion to express the left-hand side of (4.7) as 1 plus a polynomial in $(V/v) - 1$, plus a remainder. Equate to 0, solve implicitly for V and thereby show that, with $\pi(\alpha)$ defined as in (4.6),

$$\pi(\alpha) = \alpha + E(Q_\ell) + O\{(n^{-1/2}w)^{\ell+1}\} \tag{4.8}$$

for all $\ell \ge 1$, where Q_ℓ enjoys the following properties: (i) Q_ℓ is a polynomial, with constant coefficients, of degree ℓ in sums of the form $n^{-1}\sum_{i\le n} Y_i$ for random variables Y_i; (ii) if V_i is the vector whose elements are the ith summands (e.g., Y_i just above) of the respective sums in the polynomial expansion, then the V_i's are independent and identically distributed with zero mean and all moments finite; and (iii) the polynomial Q_ℓ has no term of degree 0. Note that the terms that derive from the quantity $\exp(W\Delta_\beta)$ on the left-hand side of (4.7) should explicitly preserve the factor $w = \log(1+\tau v)$, replacing W, that arises in the Taylor expansion.

The terms in Q_ℓ that are of order r for even r have expectation equal to a series in both n^{-j} and $(n^{-1/2}w)^{2j}$ for $j \ge r/2$, and the terms that are of order r for odd r have expectation equal to a similar series but for $j \ge (r+1)/2$. Arguing thus we

see from (4.8) that there exist constants $c_{j_1 j_2}$, denoting functions of θ, such that, for any $\ell \geq 2$,

$$\pi(\alpha) = \alpha + \sum_{j_1=0}^{1} \sum_{j_2=1}^{\ell-1} c_{j_1 j_2} (n^{-1} w^{2 j_1})^{j_2} + O\{(n^{-1} w^2)^{\ell}\}. \tag{4.9}$$

The value of w is asymptotic to a constant multiple of $\log m$. Therefore, result (4.9) gives (3.2) in the case $k = 1$. The constants $c_{j_1 j_2}$ are infinitely differentiable functions of the true values of the parameters β and τ, which in the first bootstrap calibration step are replaced by the maximum likelihood estimators $\hat{\beta}$ and $\hat{\tau}$. [In this step the quantity Q_ℓ in (4.8) is replaced by the analogous function of the resampled data, and the expectation in (4.8) is replaced by expectation conditional on $\mathcal{X}$.] Higher-order bootstrap iteration involves similar changes. Arguing thus we see that successive bootstrap calibration steps operate in the same way as successive bias reduction steps in more conventional problems [e.g., Hall (1992), pages 27ff.], each reducing the order of coverage error by the factor $n^{-1} w^2$.

Acknowledgments. We are grateful to two reviewers and an Associate Editor for helpful comments. In particular, the current proof of Theorem 2.1, more succinct than that of the authors, was suggested by one of the referees.

REFERENCES

BAI, C. and OLSHEN, R. A. (1988). Comment on "Theoretical comparison of bootstrap confidence intervals," by P. Hall. *Ann. Statist.* **16** 953–956.

BAI, C., BICKEL, P. J. and OLSHEN, R. A. (1990). Hyperaccuracy of bootstrap based prediction. In *Probability in Banach Spaces VII* (E. Eberlein, J. Kuelbs and M. B. Marcus, eds.) 31–42. Birkhäuser, Boston.

BARNARD, G. A. (1986). Comment on "Predictive likelihood inference with applications," by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 27–28.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.

BERAN, R. (1990). Refining bootstrap simultaneous confidence sets. *J. Amer. Statist. Assoc.* **85** 417–426.

BERAN, R. (1992). Designing bootstrap prediction regions. In *Bootstrapping and Related Techniques* (K. H. Jöckel, G. Rothe and W. Sendler, eds.) 23–30. Springer, Berlin.

BJØRNSTAD, J. F. (1990). Predictive likelihood: a review (with discussion). *Statist. Sci.* **5** 242–265.

BUTLER, R. W. (1986). Predictive likelihood inference with applications (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 1–38.

CSÖRGŐ, S., DEHEUVELS, P. and MASON, D. (1985). Kernel estimates of the tail index of a distribution. *Ann. Statist.* **13** 1050–1077.

DAVIS, R. and RESNICK, S. (1984). Tail estimates motivated by extreme value theory. *Ann. Statist.* **12** 1467–1487.

DAVISON, A. C. (1984). Modelling excesses over high thresholds, with an application. In *Statistical Extremes and Applications* (J. Tiago de Oliveira, ed.) 461–482. Reidel, Dordrecht.

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73** 323–332.

DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedences over high thresholds (with discussion). *J. Roy. Statist. Soc. Ser. B* **52** 393–442.

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events*. Springer, Berlin.

FEUERVERGER, A. and HALL, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution. *Ann. Statist.* **27** 760–781.

GRIMSHAW, S. D. (1993). Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics* **35** 185–191.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

HALL, P., PENG, L. and TAJVIDI, N. (1999). On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika* **86** 871–880.

HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174.

HOSKING, J. R. M. and WALLIS, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* **29** 339–349.

LEADBETTER, M. R. (1991). On a basis for "peaks over threshold" modeling. *Statist. Probab. Lett.* **12** 357–362.

MOHARRAM, S. H., GOSAIN, A. K. and KAPOOR, P. N. (1993). A comparative study for the estimators of the generalized Pareto distribution. *J. Hydrology* **150** 169–185.

REISS, R.-D. and THOMAS, M. (1997). *Statistical Analysis of Extreme Values, with Applications to Insurance, Finance, Hydrology and Other Fields.* Birkhäuser, Basel.

ROOTZÉN, H. and TAJVIDI, N. (1997). Extreme value statistics and wind storm losses: a case study. *Scand. Actuarial J.* 70–94.

ROSBJERG, D., MADSEN, H. and RASMUSSEN, P. F. (1992). Prediction in partial duration series with generalized Pareto-distributed exceedences. *Water Resources Research* **28** 3001–3010.

RYTGAARD, M. (1990). Estimation in the Pareto distribution. *Astin Bull.* **20** 201–216.

SMITH, R. L. (1984). Threshold methods for sample extremes. In *Statistical Extremes and Applications* (J. Tiago de Oliveira, ed.) 621–638. Reidel, Dordrecht.

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90.

SMITH, R. L. (1987). Estimating tails of probability distributions. *Ann. Statist.* **15** 1174–1207.

SMITH, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone (with discussion). *Statist. Sci.* **4** 367–393.

STINE, R. A. (1985). Bootstrap prediction intervals for regression. *J. Amer. Statist. Assoc.* **80** 1026–1031.

ZIPF, G. K. (1941). *National Unity and Disunity: The Nation as a Bio-Social Organism*. Principia Press, Bloomington, IN.

ZIPF, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, MA.

P. HALL
CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA
E-MAIL: halpstat@pretty.anu.edu.au

L. PENG
CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA

N. TAJVIDI
CENTRE FOR MATHEMATICAL SCIENCES
LUND INSTITUTE OF TECHNOLOGY
SE-22100 LUND
SWEDEN