Inference for extreme values under threshold-based stopping rules

Anna Maria Barlow, Chris Sherlock, Jonathan Tawn

STOR-i Centre for Doctoral Training, Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, U.K.

Abstract

There is a propensity for an extreme value analyses to be conducted as a consequence of the occurrence of a large flooding event. This timing of the analysis introduces bias and poor coverage probabilities into the associated risk assessments and leads subsequently to inefficient flood protection schemes. We explore these problems through studying stochastic stopping criteria and propose new likelihood-based inferences that mitigate against these difficulties. Our methods are illustrated through the analysis of the river Lune, following it experiencing the UK's largest ever measured flow event in 2015. We show that without accounting for this stopping feature there would be substantial over-design in response to the event.

Keywords: Extreme value theory, flooding, stopping rules

1 Introduction

The UK currently spends £400-500M per year on coastal and river flood defence infrastructure, with 2 million properties exposed to the risk of flooding. The agencies responsible for this spend monitor the effectiveness of their investment at giving the level of protection expected. After major flooding events renewed analysis is performed to assess both existing flood defences and the cost benefit of potential new schemes, proposed in response to the flooding.

Statistical extreme value methods, with likelihood-based inference, have proved a core component of the required analysis in terms of minimising the costs without jeopardising the level of accepted risk, and hence have financial and societal benefits. However, there is a problem with using these methods when the statistical analysis has been prompted by the occurrence of a recent large event, since in this case the data-set size itself is also random. This can lead to substantially biased inference and poor coverage properties and so result in inefficient flood-defence designs. Omitting the new extreme data value from the data set also seems unsuitable, as intuition suggests that flood risk will then be underestimated; moreover it would appear perverse to flood management agencies to ignore events of the type most relevant to the design specification.

This paper aims to identify the extent of the inference problems when an analysis has been triggered by a large event and to develop new conditional-likelihood methods which appear to overcome these problems. We do not *suggest* when the timing of the data analysis should take place but study the analysis given that its timing has been determined by a data-dependent decision making process.

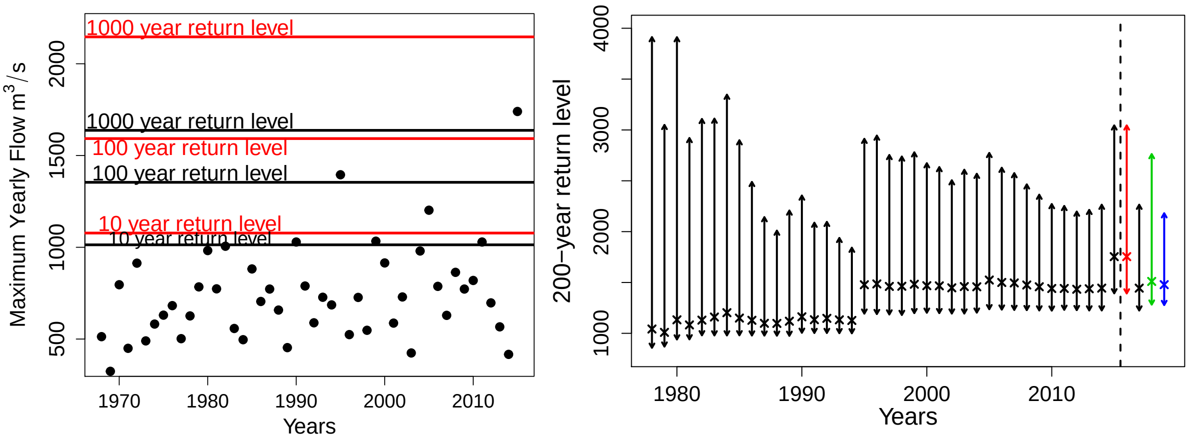

Figure 1: *Left: The annual maxima of daily peak river flow data for the Lune at Caton with return level estimates before (black) and after (red) the 2015 flood. Right: 200-year return-level estimates based on all the data up to and including the current year for the Lune at Caton with 95% profile likelihood-based confidence intervals. The four 200-year return-level estimates and associated 95% confidence intervals to the right of the vertical dotted line are our new estimates that aim to address a fixed-threshold stopping rule of* $c_k = 1568$ *based on the all the data up to and including 2015: standard likelihood (red), excluding the final observation (black), full conditioning (green) and partial conditioning (blue).*

We consider modelling the extreme events of a time series of independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$. The classical approach to do this is to split the time series into blocks of equal size (often a year) and to model the maxima of these blocks. Linear normalisation is necessary since as the block size tends to infinity the distribution of the maxima degenerates to a point mass at the upper end point of the distribution of X. The generalised extreme value (GEV) distribution (Coles, 2001) is the only non-degenerate limiting distribution of the normalised maxima as the block size tends to infinity. The GEV has distribution function:

$$G(x) = \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right), \tag{1.1}$$

where μ, $\sigma > 0$ and ξ are the location, scale and shape parameters respectively and $[z]_+ = \max(z, 0)$. The shape parameter determines the behaviour of the upper tail of the distribution: for $\xi < 0$ the distribution has an upper end point, for $\xi = 0$ the tail is exponential and for $\xi > 0$ the distribution has a power-decaying tail. There is particular interest in the occurrence of extreme events and so an important part of the analysis is the estimation of return-levels (quantiles). Under stationarity, the y-year return-level, x_y, is the value which is exceeded on average once every y years. For the GEV distribution this can be calculated as:

$$x_y = G^{-1}\left(1-\frac{1}{y}\right) = \begin{cases} \mu - \frac{\sigma}{\xi}\left\{1-\left[-\log\left(1-1/y\right)\right]^{-\xi}\right\} & \xi \neq 0 \\ \mu - \sigma\log\left[-\log\left(1-1/y\right)\right] & \xi = 0. \end{cases} \tag{1.2}$$

One can also consider modelling daily observations above some high threshold (rather than just modelling the block maxima) by the asymptotically justified generalised Pareto distribution (GPD) (Davison and Smith, 1990). Threshold methods typically benefit from using more extreme-value data and hence are more efficient in their inferences than block maxima methods (Coles, 2001). We focus most of our analysis and developments on the GEV case, as similar benefits are found for both GEV and GPD inference, but with the GPD also sensitive to threshold choice. We illustrate some GPD results in §S.2 of the supplementary material.

We consider the analysis of annual maxima of daily peak river flow data obtained from CEH (2018) for the Lune at Caton, just outside Lancaster, from 1968 to 2015 (Figure 1, left panel) and illustrate the inference issues due to the timing of analysis being determined by the occurrence of a flood event. Under the assumption that the annual maxima are i.i.d. we can fit the GEV distribution to the annual maxima using likelihood-based inference (the likelihood is $\prod_{i=1}^{n} g(x_i; (\mu, \sigma, \xi))$ where n is the sample size and g is the density, $g = dG/dx$), and estimate return-levels using (1.2). The estimated 10, 100 and 1000-year return levels are shown in Figure 1 (left panel) for the data up to 2014 (black) and including 2015 (red). The December 2015 floods resulted in the river Lune recording the highest peak river flow (1740 m^3/s) of all UK rivers over all years of records. This value is higher than the 1000-year return-level estimate based on the observations up to 2014. However, once the 2015 event is included in the analysis the return-level estimates become much higher. If we were to take these 2015 point estimates as the truth we would expect to observe an event as extreme as that in 2015 approximately once every 200 years. For design purposes this level of sensitivity is highly undesirable, as the costs for flood protection would change dramatically.

Figure 1 (right panel) shows a reanalysis of all data available at each year between 1978 and 2015. It provides the point estimate and profile likelihood-based 95% confidence interval of the 200-year return-level, as it would have been produced in that year. The four additional point estimates and confidence intervals to the right of the vertical dotted line correspond to estimators introduced in §3 and their corresponding profile likelihood-based confidence intervals. At the beginning of the data collection the return-level estimates vary considerably, but they become more stable as the number of years increases, with the width of the confidence intervals generally decreasing over time. However, even after many years of data collection, the largest events can be seen to cause sharp increases in the estimates and their associated uncertainty. For example, the return-level estimate following the January 1995 floods and the 2015 floods are larger than those of previous years.

This illustrative example is typical of when an analysis is performed immediately after a large event. Unless further analysis is undertaken it is unclear whether by analysing the data with the final extreme event we are introducing a positive bias into the inference. For example, the lower bound of the 95% confidence interval of the 200 year return-level after the 1995 event is larger than almost all previous point estimates - directly after the event (without the knowledge of later years) this could have been seen as an indication of positive bias in the standard estimator. However, after the 1995 event the return-level estimates were fairly stable and larger than those before 1995, so it would seem the standard estimator for 1995 may not have been overestimating and before the event the shape parameter was estimated too low.

An alternative approach is to simply ignore the most recent year of data when an analysis has been requested because we have large data in that year, in which case the return-level estimate is lower and the confidence interval is narrower - in particular the upper bound is lower. However, we speculate (see also §2.3) that this estimator is now negatively biased due to the loss of information about the extreme event. Moreover the estimator is inefficient since the larger data values are the most informative about the upper tail (Davison and Smith, 1990). Finally, it would be hard to convince practitioners to exclude the largest events; for example, an event may be observed which is larger than the upper end point estimated from previous data, in which case it would be perverse not to make some update to the previously estimated return-levels.

The key issue that the Lune example illustrates is that when meeting the flood management agencies' needs, the time to undertake the extreme value analysis is stochastic and triggered by a large event. Thus, there is effectively some form of unwritten stopping rule, determined by the flood management agencies, which determines the timing of the analysis. In contrast with a standard i.i.d. sample of fixed size, when we use a stopping rule the time at which we stop (the sample size) is variable, we denote this by N.

One can attempt to mathematically formulate the characteristics of the stopping rule, though in reality a precise mathematical rule does not exist. The stopping decision may depend on (i) some absolute threshold, such as the height of existing flood defences or a critical level which when exceeded leads to severe flooding, or (ii) an assessment, based upon all observations to date, of what might constitute an 'exceptional' event. We consider two simple stopping rules based on a series of i.i.d. random variables, $X_1, X_2, \ldots$ which, in a sense, bracket this range of possibilities and we discuss other possibilities in §6.

1. **Fixed-threshold stopping rule**
   Stop when an observation exceeds a specified value, c_k, *i.e.*:
   $$N = \inf\{n \in \mathbb{N} : X_n > c_k\}\,, \tag{1.3}$$
   where k is the true (but unknown) return period of c_k.
2. **Variable-threshold stopping rule**
   Stop when an observation exceeds the return-level estimate, $\hat{x}_k$, corresponding to the fixed return period of k years, calculated using previous observed values, *i.e.*, when:
   $$N = \inf\{n \in \mathbb{N} : X_n > \hat{x}_k(X_1, \ldots, X_{n-1})\}\,. \tag{1.4}$$

We do not suggest the stopping rule to use but study the analysis given that its timing has been determined by a stopping rule. As far as we are aware, there has been no study of stopping rules and their effects on likelihood estimation in the extreme-value setting.

Using a stopping rule to determine the sample size can lead to estimators, such as the maximum likelihood estimator (MLE), having different sampling properties to the fixed-sample case. To illustrate this feature consider an i.i.d. sequence of Bernoulli random variables, $Y_1, Y_2, \ldots$, each with probability of success of θ. If one fixes the number of trials, n, the number of successes, R, in these trials is binomially distributed and $\hat{\theta} = R/n =: \hat{\theta}_1$; whereas if the number of successes is fixed as r, the number of trials, N, is negative-binomially distributed and $\hat{\theta} = r/N =: \hat{\theta}_2$. In both cases the MLE of the probability of success is the proportion of successes, however, $\mathbb{E}\left[\hat{\theta}_1\right] = \theta$ whereas $\mathbb{E}\left[\hat{\theta}_2\right] = r\mathbb{E}\left[1/N\right] \geq r/\mathbb{E}\left[N\right] = \theta$ by Jensen's inequality. The presence of a stopping rule affects the performance of the estimator which motivates an investigation into the performance of return-level estimation under stopping rules.

Testing the data against some 'stopping criterion' at regular intervals falls into the setting of sequential analysis, which has a rich literature covering applications from quality control (Wald, 2004), to clinical trials (Todd et al., 1996) and abundance modelling (Barry and Coggan, 2010). Many studies have considered the influence of such stopping rules on likelihood inference, *e.g.*, Barndorff-Nielsen and Cox (1984) consider the distribution of the likelihood-ratio statistic under different stopping rules for systems with Brownian motion and Poisson processes, and in the clinical trial setting Whitehead (1986) derives an expression for the bias of the MLE of the treatment effect tested under a sequential probability ratio test. Some papers compare the bias under different experimental designs or stopping criteria (*e.g.*, Bauer et al. (2010)). Cox (1952), Whitehead (1986) and Stallard and Todd (2005) propose bias-reduced estimators by approximating the bias and subtracting this from the usual estimate. One such approach uses an iterative method corresponding to a bootstrap bias correction (Efron, 1990).

Kenward and Molenberghs (1998) consider i.i.d. sampling from a Normal distribution using a deterministic stopping rule and study the estimation of the mean parameter of this Normal distribution. Molenberghs et al. (2014) extend this setting to the use of a probabilistic stopping rule. They note that an unbiased estimator of the mean parameter can be obtained from the conditional likelihood (we derive such estimators in §3) however, at the cost of an increased mean squared error (MSE) in comparison to the MSE of the sample average (the standard estimator if the sample size was fixed). The increased variance of a bias-reduced estimator appears to be an issue for many of the proposed bias reduction methods. For example, bias reduction using Rao-Blackwellisation (Bowden and Glimm, 2008) and shrinkage estimators (Carreras and Brannath, 2013) often have a worse MSE than the standard MLEs.

In §2 we introduce the notation used throughout the paper and discuss likelihood inference under stopping rules. In §2.3 and §2.4 we discuss the bias under the fixed-threshold and variable-threshold stopping rules respectively and derive expressions for the bias when the data arise from an exponential distribution. We introduce two conditioning-based likelihood estimators in §3. In §4 we perform a simulation study for sampling from the GEV distribution using the two stopping rules and discuss the properties of the estimators in this setting. We apply our estimators to the Lune river flow data in §5 and discuss our conclusions, the practical usage of the methods and extensions in §6.

2 Inference under stopping rules

2.1 Introduction

Throughout this paper we restrict our attention to sequences of iid observations arising from some distribution with a density of $f(x;\theta)$, where θ is the parameter vector for the distribution. When a fixed number, n, of observations, $x_1,\ldots,x_n$, is analysed, the likelihood and log-likelihood for the data are

$$L_{fixed}(\theta;\boldsymbol{x}_{1:n})=\prod_{i=1}^{n}f(x_i;\theta) \quad\text{and}\quad \ell_{fixed}(\theta;\boldsymbol{x}_{1:n})=\sum_{i=1}^{n}\log f(x_i;\theta),$$

where $\boldsymbol{x}_{1:n}$ denotes the vector of observations $(x_1,\ldots,x_n)$. In practice it is usual to assume the sample size, n, is fixed however, for us, n, is not fixed; we sample consecutively until some stopping criterion is met and denote the (random) time at which it is met by N. In Section 2.2 we detail a well-known result that the likelihood for the data $(n, x_{1:n})$ with a random stopping time is the same as for the data $x_{1:n}$ with n fixed, except for a constant of proportionality. In this article we consider both the full data and excluding the last data point, for which the log-likelihoods, respectively denoted by ℓ_{std} and ℓ_{ex}, are, up to an additive constant:

$$\begin{aligned}\ell_{std}(\theta;n,\boldsymbol{x}_{1:n}) &= \ell_{fixed}(\theta;\boldsymbol{x}_{1:n}) && (2.1)\\ \ell_{ex}(\theta;n,\boldsymbol{x}_{1:n}) &= \ell_{std}(\theta;n,\boldsymbol{x}_{1:n})-\log f(x_n;\theta)\,. && (2.2)\end{aligned}$$

Given data $(n,\boldsymbol{x}_{1:n})$ an estimate of the parameter vector is obtained by maximising the log likelihood: $\hat{\theta}(n,\boldsymbol{x}_{1:n})=\arg\max_\theta \ell(\theta)$. When the nature of the data is clear we abbreviate this to $\widehat{\theta}$, and depending on the likelihood used we have estimators $\hat{\theta}_{std}$ or $\hat{\theta}_{ex}$. Although inference based purely on ℓ_{std} gives the same result as inference based purely on ℓ_{fixed} the properties of the resulting estimators, such as bias and variance, as well as the coverage of any confidence interval, may be influenced by the different data-generating mechanisms.

In practice we would not consider estimating return-levels (particularly for large return periods) from a sample of only a very small number of observations. However, the fixed-threshold stopping rules can result in samples of size 1, and this can lead to parameter identifiability issues for data sets simulated from the hypothesised data-generating mechanism. In reality, if an analysis has been requested then sufficient information would be available to derive a meaningful estimate. This information could be historical information, hydrological knowledge, data from other sites, or data at the current site collected before the instigation of a stopping rule. We call this the *historical data* and, for simplicity in this article, code the historical data as some number, n_0 of data values collected before the stopping rule could be invoked. Real decisions will incorporate this information, and our analysis should allow for this.

We are interested in a set of y-year return-levels, $x_y(\theta)$ $(y \in \mathcal{Y})$, for some set $\mathcal{Y}$, such as $\{50, 200, 1000\}$. In particular, we wish to understand the behaviour of the estimators $x_y(\widehat{\theta}(N, \boldsymbol{X}_{1:N}))$ (with x_y given by expression (1.2) for GEV sampling) when the dataset arises from a stopping rule. In this section we focus on the relative bias, and in §4 we look at other properties including the relative root-mean-squared error, given respectively by:

$$\text{RelBias}(\hat{x}_y) = \frac{1}{x_y(\theta)} \mathbb{E}\left[x_y(\widehat{\theta}(N, \boldsymbol{X}_{1:N}))\right] - 1 \tag{2.3}$$

$$\text{RRMSE}(\hat{x}_y) = \frac{1}{x_y(\theta)} \sqrt{\mathbb{E}\left[\{x_y(\widehat{\theta}(N, \boldsymbol{X}_{1:N})) - x_y(\theta)\}^2\right]}, \tag{2.4}$$

where $x_y(\theta)$ is the true y-year return-level.

The properties of likelihood-based estimators of tail quantiles under our stopping rules are intractable for data arising from the GEV or GPD distributions. However, for a particular special case of the GPD, the exponential distribution, certain properties are tractable and this provides insight into the behaviour observed in the simulation studies of §4 for the GEV. Specifically for exponential data, in §2.3 we derive the bias in quantile estimates under the fixed-threshold stopping rule, and in §2.4 show that, under the variable-threshold stopping rule, quantile estimates are unbiased.

2.2 Likelihood in presence of a stopping rule

Now, following Pawitan (2013), we derive the true likelihood for the data sampled using a general stopping rule which is a function of the data and not the unknown parameter vector. We define a stopping region $\mathcal{S}_n = \mathcal{S}_n(\boldsymbol{x}_{1:n-1})$ such that we stop sampling if $X_n \in \mathcal{S}_n$ and continue to sample otherwise. We abbreviate $\mathbb{P}(X_i \in \mathcal{S}_i)$ by p_i and we let $f_{X_i|\mathcal{S}_i}$ and $f_{X_i|\mathcal{S}_i^c}$ denote the densities of X_i conditional on $X_i \in \mathcal{S}_i$ and $X_i \in \mathcal{S}_i^c$. The likelihood for the full data is

$$\begin{aligned} L_{std}(\theta; n, \boldsymbol{x}_{1:n}) &= \mathbb{P}(N = n, \boldsymbol{X}_{1:n} = \boldsymbol{x}_{1:n} | \theta) \\ &= \mathbb{P}(N = n)\, \mathbb{P}(\boldsymbol{X}_{1:n} = \boldsymbol{x}_{1:n} | N = n, \theta) \\ &= p_n \mathbb{1}_{\mathcal{S}_n}(x_n) \prod_{i=1}^{n-1} (1 - p_i) \mathbb{1}_{\mathcal{S}_i^c}(x_i) \times f_{X_n|\mathcal{S}_n}(x_n|\mathcal{S}_n) \prod_{i=1}^{n-1} f_{X_i|\mathcal{S}_i^c}(x_i|\mathcal{S}_i) \qquad (2.5) \\ &= L_{fixed}(\theta; \boldsymbol{x}_{1:n}) \times \mathbb{1}_{\mathcal{S}_n}(x_n) \prod_{i=1}^{n-1} \mathbb{1}_{\mathcal{S}_i^c}(x_i) \\ &\propto L_{fixed}(\theta; \boldsymbol{x}_{1:n}). \end{aligned}$$

The logic here is that to have a sample of size n the final observation must be in the stopping region and all other observations outside their respective stopping regions, hence $\mathbb{P}(N = n)$ includes indicator functions of the observations being in the correct sets. The last step follows since the indicator functions do not depend on the unknown parameter, θ, and so are absorbed into the proportionality constant. Thus inference purely from the likelihood leads to the same conclusions whether we have a random sample size according to some stopping rule or a fixed sample size. In particular, the MLE, $\widehat{\theta}$, and the observed Fisher information are the same in both cases. However, the properties of the estimators are different since the distribution of $\{N, X_1, \ldots, X_N\}$ is different to the distribution of $\{X_1, \ldots, X_n\}$ for some fixed n. In particular, estimators obtained from L_{std} can be biased even when estimators from L_{fixed} are unbiased, as seen in §1 for Bernoulli sampling.

2.3 Fixed-threshold stopping rule with exponential observations

Let X_i have an exponential distribution with an unknown rate parameter of β. This distribution is a special case of the GPD, which used to model the tails of a distribution, and is given by expression (S.1) in the Supplementary Material when $\xi = 0$ and $\sigma = \beta^{-1}$. The y-observation return-level is $x_y = (\log y)/\beta$ and, since this is proportional to $1/\beta$, the relative bias is $\beta/\widehat{\beta} - 1$ whatever the value of y. The MLE of β^{-1} for a sample of size n, whether fixed or random is simply $\bar{x}$, where $\bar{x}$ is the sample mean. When n is fixed, the MLE, $(\widehat{\beta}_{fixed})^{-1} = \overline{X}_n$, is an unbiased estimator of $1/\beta$; however with the fixed-threshold stopping rule N follows a geometric distribution where $1/k$ is the probability of a 'success' *i.e.*, an exceedance. The geometric distribution is a special case of the negative-binomial distribution and so we know the estimator of the probability of exceedance of a fixed threshold is positively biased (§1 under (1.4)). Now $(\widehat{\beta}_{std})^{-1} = \overline{X}_N$ and, similarly, the MLE when excluding the final observation is $(\widehat{\beta}_{ex})^{-1} = \bar{X}_{N-1} = \sum_{i=1}^{N-1} X_i/(N-1)$. It is straightforward to show (see Appendix A) the following.

Proposition 1 *Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $X_i \sim \mathsf{Exp}(\beta)$. Let N arise from the fixed-threshold stopping rule (1.3) giving data $(N; \boldsymbol{X}_{1:N})$. Let $\hat{x}_y^{std} = x_y(\widehat{\beta}_{std}(N; \boldsymbol{X}_{1:N}))$ be the estimator of the $\left(1 - \frac{1}{y}\right)$th quantile (equivalently the y-observation return-level) obtained from the MLE for β from the full likelihood and let $\hat{x}_y^{ex} = x_y(\widehat{\beta}_{ex}(N; \boldsymbol{X}_{1:N}))$ be the estimator from the likelihood excluding the final observation.*

Then the relative biases of the return-level estimators are

$$\begin{aligned}\frac{1}{x_y}\mathbb{E}\left[\hat{x}_y^{std}\right] - 1 &= \frac{\beta c_k}{e^{\beta c_k} - 1}\left(\frac{\beta c_k}{1 - e^{-\beta c_k}} - 1\right)\\ \frac{1}{x_y}\mathbb{E}\left[\hat{x}_y^{ex} | N > 1\right] - 1 &= -\frac{\beta c_k}{e^{\beta c_k} - 1}.\end{aligned}$$

In Proposition 1, when excluding both the final observation (and the fact that it *is* the final observation), when $N = 1$ the MLE is undefined since there are no data; x_1 is unknown and the fact that N would be greater than zero was known before the data-collection process began; we therefore condition on $N > 1$.

Proposition 1 shows that the estimator of any return-level using the full likelihood is always positively biased, whereas if the final observation is omitted the estimator of any return-level is always negatively biased. The final data observation is the largest and has been shown by Davison and Smith (1990) to be the most influential on the MLE fit so when this value, together with the information that it exceeded the threshold, is omitted from the dataset this changes the bias and, potentially, also the variance of the return-level estimator and risks being inefficient. Nevertheless, for thresholds with only a small chance of exceedance, *i.e.*, large values of βc_k, $\text{RelBias}(\hat{x}_y^{std}) \sim (\beta c_k)^2 \exp(-\beta c_k)$, whereas $\text{RelBias}(\hat{x}_y^{ex}) \sim -\beta c_k \exp(-\beta c_k)$, that is, the bias is a factor $(\beta c_k)^{-1}$ smaller for estimates where the final observation is ignored. The higher the threshold, the larger the typical data set that is generated before the stopping criterion is met and the less biased the estimate of any return-level.

In Figure 2 we compare the relative bias of the estimates of β^{-1} (and hence also for the return-level estimates) both when including and excluding the final observation when varying k, the

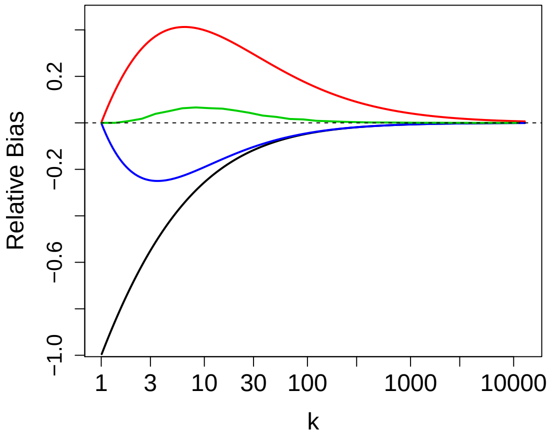

Figure 2: *Relative bias of the return-level estimates against the return period, k, of the fixed-threshold stopping rule c_k when sampling from the Exponential distribution with the fixed threshold stopping rule using: standard likelihood (red), excluding the final observation (black), full conditioning (green) and partial conditioning (blue).*

true return period of the stopping threshold, c_k. The two additional curves correspond to estimators that will be introduced in §3. The maximum relative bias in the standard return-level estimator is 0.4; *i.e.*, the estimator is around 1.4 times the true value. This occurs for a threshold corresponding to $k \approx 7$, *i.e.*, when we stop sampling if an observation exceeds the 7-observation return-level. Clearly this will generally result in a very small sample so we would expect return-level estimates to also be highly variable in this case.

2.4 Variable-threshold stopping rule with gamma observations

The positive bias in return-level estimates that arises from the fixed stopping rule is partly a result of the geometric distribution of N (see §2.3). For the variable-threshold rule N no longer has a geometric distribution and we find empirically for the GEV (see §4.3) that the bias is typically reduced; as we now show, at least for one parametric family of distributions, the bias disappears entirely.

Let $X_i \sim \mathsf{Exp}(\beta)$, with unknown rate parameter, $\beta > 0$. In §2.1 we noted the need for a historical sample in practice; here, to reflect this, we suppose that the stopping rule is only implemented after an initial sample of independent $\mathsf{Exp}(\beta)$ variables, $X_{-n_0}, \ldots, X_{-1}$, whose mean is denoted by $\bar{X}_0$, with $\bar{X}_0 \sim \mathsf{Gamma}(n_0, n_0\beta)$.

As noted earlier, the return level is proportional to β^{-1} and the MLE for β^{-1}, from the full likelihood, is $\bar{x}$. Thus, for some constant of proportionality γ (depending the return period, y), the variable-threshold stopping rule is equivalent to

$$N = \inf\{n \geq 1 : X_n > \gamma \overline{X}_{n-1}\}, \tag{2.6}$$

where

$$\overline{X}_k = \frac{n_0\overline{X}_0 + X_1 + \ldots + X_k}{n_0 + k} \qquad k \geq 1\,. \tag{2.7}$$

Theorem 1 *With N, $\boldsymbol{X}_{1:N}$ and $\overline{X}_k$ as defined in (2.6) and (2.7), for all $n \in \mathbb{N}$:*

$$\overline{X}_N | N = n \quad \overset{d}{=} \quad \overline{X}_n \sim \mathsf{Gamma}((n + n_0), \beta(n + n_0))\,.$$

From Theorem 1 we see that $\mathbb{E}\left[\overline{X}_n | N = n\right] = 1/\beta$, and hence:

Corollary 1 *For a sample obtained as in Theorem 1, the sample mean and y-year return-level estimate are unbiased:*

$$\mathbb{E}\left[\overline{X}_N\right] = \frac{1}{\beta}.$$
$$\mathbb{E}\left[x_y(\widehat{\beta}^{std}(N, \boldsymbol{X}_{1:N}))\right] = x_y(\beta).$$

Contrasting Corollary 1 with Proposition 1, both of which apply to the exponential distribution, we see that the standard estimator can be unbiased for the variable-threshold stopping rule even though it is strongly positively biased for the fixed-threshold stopping rule.

Theorem 1 and Corollary 1 can be extended to $X_i \sim \mathsf{Gamma}(\alpha, \beta)$ random variables, where the shape parameter, $\alpha > 0$, is known and the the rate parameter, $\beta > 0$, is unknown (and must be estimated). As with the exponential distribution, the return-levels of the gamma distribution are proportional to β^{-1}, with the constant of proportionality depending on α. Furthermore the MLE from the full likelihood satisfies $\widehat{\beta^{-1}} = \bar{x}/\alpha$. So, the variable-threshold stopping rule is (2.6) with the constant of proportionality γ depending on α as well as the return period, y. Theorem 1 and Corollary 1 are retrieved by setting $\alpha = 1$.

A proof for Theorem 1 in this more general case is provided in Appendix A.2.

3 Alternative Methods for Parameter Inference

3.1 New conditional likelihoods

Motivated by the lack of bias in the stopping rule of Molenberghs et al. (2014), we propose a similar estimator for our scenarios by conditioning on the fact that only the final observation met the stopping criterion. The likelihood, therefore, consists of the conditional densities of the data values given that each of the first $n-1$ is outside its stopping region and the nth is inside its stopping region. Using the same notation as §2.2 the full conditioning likelihood is, from (2.5),

$$\begin{aligned} L_{fc}(\theta; n, \boldsymbol{x}_{1:n}) &= \frac{\mathbb{P}\left(N = n, \boldsymbol{X}_{1:n} = \boldsymbol{x}_{1:n} | \theta\right)}{\mathbb{P}\left(X_n \in \mathcal{S}_n\right) \prod_{i=1}^{n-1} \mathbb{P}\left(X_i \in \mathcal{S}_i^c\right)} \\ &= f_{X_n|\mathcal{S}_n}(x_n|\mathcal{S}_n) \prod_{i=1}^{n-1} f_{X_i|\mathcal{S}_i^c}(x_i|\mathcal{S}_i) \times \mathbb{1}_{\mathcal{S}_n}(x_n) \prod_{i=1}^{n-1} \mathbb{1}_{\mathcal{S}_i^c}(x_i). \end{aligned}$$

The log likelihood, ℓ_{fc}, is as given in (3.1) and the corresponding estimate is denoted $\hat{\theta}_{fc}$. For the fixed-threshold stopping rule this effectively conditions out the geometric distribution for N (§2.3); it might be hoped, therefore, that it might remove that part of the positive bias that is due to the randomness of N.

By conditioning on the final observation exceeding its stopping threshold and all other observations not exceeding theirs we are effectively losing all of this information which will lead to larger uncertainty in the estimates, *e.g.*, giving wider confidence intervals. Hence, we consider a further likelihood which conditions *only* on the fact that the final observation exceeds its threshold:

$$\frac{\mathbb{P}\left(N = n, \boldsymbol{X}_{1:n} = \boldsymbol{x}_{1:n} | N = n, \theta\right)}{\mathbb{P}\left(X_n \in \mathcal{S}_n\right)} = f_{X_n|\mathcal{S}_n}(x_n|\mathcal{S}_n) \prod_{i=1}^{n-1} f_{X_i}(x_i) \times \mathbb{1}_{\mathcal{S}_n}(x_n) \prod_{i=1}^{n-1} \mathbb{1}_{\mathcal{S}_i^c}(x_i).$$

As with full conditioning, this results in the stochasticity of N being less influential. We refer to this method as partial conditioning with log likelihood, denoted by ℓ_{pc}, given in (3.2). The corresponding estimate is denoted by $\hat{\theta}_{pc}$.

In summary, the two new log likelihoods we consider are:

$$\begin{aligned} \ell_{fc}(\theta; n, \boldsymbol{x}_{1:n}) &= \ell_{std}(\theta; n, \boldsymbol{x}_{1:n}) - \log \bar{F}(s_{k,n}; \theta) - \sum_{i=1}^{n-1} \log F(s_{k,i}; \theta) && (3.1) \\ \ell_{pc}(\theta; n, \boldsymbol{x}_{1:n}) &= \ell_{std}(\theta; n, \boldsymbol{x}_{1:n}) - \log \bar{F}(s_{k,n}; \theta) && (3.2) \end{aligned}$$

where $s_{k,i}$ is the lower boundary of the stopping set for the ith observation; then for the variable-threshold stopping rule $s_{k,i} = \hat{x}_k^{std}(\boldsymbol{x}_{1:i-1})$, *i.e.*, it is the standard estimate of the k-year return-level using all the data up to and including the previous observation, and for the fixed-threshold stopping rule $s_{k,i} = c_k$ for all i.

The examples of §2 have exponential tails with an unknown scale parameter. When data values are modelled using the GEV, uncertainty in the shape parameter, ξ, has a much larger impact on estimates of high quantiles than the uncertainty in the other two parameters, μ and σ (Coles, 2001). So we now consider estimation of the shape parameter and high quantiles using the standard and partial conditioning likelihoods. For simplicity we focus on an idealised scenario where we take $\mu = 0$ and $\sigma = 1$ as known, so X has a distribution function of $F(x; \xi) = \exp\left(-[1 + \xi x]_+^{-1/\xi}\right)$ and a survival function of $\bar{F} = 1 - F$.

For quantile estimation the standard likelihood estimator of ξ, *i.e.*, $\hat{\xi}_{std}$, leads to a positive bias for high quantiles. This can be seen as follows. The y-year return level can be written as $\bar{F}^{-1}(1/y; \xi) = [\exp(a_y \xi) - 1]/\xi$, with $a_y = -\log[-\log(1 - 1/y)]$, where $a_y \geq 0$ provided $y \geq e/(e-1) \approx 1.6$. Return levels as low as 1.6 years are of no practical interest in our setting. When $a_y > 0$, $\bar{F}^{-1}$ is an increasing, convex function of $\xi \in \mathbb{R}$, so, whatever the likelihood, Jensen's inequality gives $\mathbb{E}_\xi\left[\bar{F}^{-1}(1/y; \xi)\right] \geq \bar{F}^{-1}(1/y; \mathbb{E}_\xi[\xi])$. The monotonicity of $\bar{F}^{-1}$ implies that even if $\hat{\xi}_{std}$ is unbiased, we should expect a positive bias in all quantile estimates, and this will only be exaggerated if (as we find in our stopping-rule simulations) ξ is positively biased.

This bias in the estimator for high quantiles is guaranteed to be less positive when using the partial conditioning likelihood rather than the standard likelihood. To see this first note that $\ell_{pc}(\xi) = \ell_{std}(\xi) - \log[\bar{F}(c; \xi)]$ where c, the stopping threshold, has been standardised. The resulting MLE, $\hat{\xi}_{pc}$, satisfies $\ell(\hat{\xi}_{pc}) - \log[\bar{F}(c; \hat{\xi}_{pc})] > \ell(\xi) - \log[\bar{F}(c; \xi)]\ \forall\ \xi$. Also, $\bar{F}$ is an increasing function of ξ since

$$\frac{\partial}{\partial \xi} \log\{-\log F(x; \xi)\} = \frac{1}{\xi^2}\left\{\log[1 + \xi x] - \frac{\xi x}{1 + \xi x}\right\} \geq 0, \qquad \forall\ \xi x > -1\,.$$

So, as $\ell(\hat{\xi}_{std}) > \ell(\xi)\ \forall\ \xi$, it follows that

$$\begin{aligned} &\ell(\hat{\xi}_{std}) - \log[\bar{F}(c; \hat{\xi}_{pc})] > \ell(\hat{\xi}_{pc}) - \log[\bar{F}(c; \hat{\xi}_{pc})] > \ell(\hat{\xi}_{std}) - \log[\bar{F}(c; \hat{\xi}_{std})] \\ &\qquad \Rightarrow -\log[\bar{F}(c; \hat{\xi}_{pc})] > -\log[\bar{F}(c; \hat{\xi}_{std})] \\ &\qquad\qquad \Rightarrow \hat{\xi}_{pc} < \hat{\xi}_{std}. \end{aligned}$$

Thus, if the standard estimator is positively biased the partial conditioning method will be less positively biased. Given that this effect is magnified for return levels, as shown above, we should expect improvements in return level estimates using the partial conditioning likelihood. An analogous argument to the above also applies to data modelled using the GPD, except that there is no restriction on y, and $a_y = \log y$.

3.2 Application to exponential observations

Consider iid sampling from the exponential distribution with rate parameter, β, using the fixed-threshold stopping rule. The relative bias for return-level estimators using the log-likelihoods (2.1) and (2.2), are detailed in Proposition 1 and plotted in Figure 2. Figure 2 also plots the relative bias for the likelihoods in (3.1) and (3.2), the latter has the form

$$\text{RelBias}(\hat{x}_y^{pc}) = \text{RelBias}(\hat{x}_y^{std}) - \frac{\beta^2 c_k^2}{e^{\beta c_k} - 1} = \frac{\beta c_k}{e^{\beta c_k} - 1}\left[\frac{\beta c_k}{1 - e^{-\beta c_k}} - 1 - \beta c_k\right]. \qquad (3.3)$$

Estimator $\hat{x}_y^{pc}$ is negatively biased but the bias is smaller than that for $\hat{x}_y^{ex}$. The bias of $\hat{x}_y^{pc}$ tends to 0 as c_k tends to infinity at the same fast rate as for $\hat{x}_y^{ex}$ (§2.3).

We were unable to obtain a tractable expression for the bias of the full-conditional estimator. In Figure 2 this bias was found using Monte Carlo methods. The bias is very low and tends towards 0 much faster than any of the other estimators considered. This finding is similar to that of Molenberghs et al. (2014) for the mean of normally distributed observations with a probabilistic stopping rule; however, the MSE of the unbiased estimator was found to be poor compared to that of the standard estimator. In §4 we show that in our 'extremes' setting, the full-conditional MSE for a return-level is often lower relative to the MSE of the standard estimator since the high variance of return-level estimators using the standard likelihood is in part due to the final observation being large. Furthermore in §4 we show that the partial conditioning approach results in estimators with much reduced variance and that this leads to lower MSE compared to the standard likelihood approach.

4 Simulation results

In this section we focus on the return-level inference when sampling from the GEV distribution with the two stopping rules of §1. In §4.2 we calculate the fixed stopping threshold, c_k, for a range of return periods, k, using (1.2) and our knowledge of the true parameters μ, σ and ξ. In §4.3 we consider the variable threshold stopping rule over a range of k. Similar simulation results are given in the supplementary material for the GPD.

4.1 Simulation design

We investigate true return-periods, k, between 20 and 2000. When generating the data, for each k, for the fixed-threshold stopping rule, we set c_k to be the true $(1 - 1/k)$th quantile of the data-generating distribution (*i.e.*, the k-yr return-level) whereas for the variable-threshold rule the threshold is the estimated $(1 - 1/k)$th quantile; with both rules we stop at the first exceedance. In the simulation study, for each combination of θ, stopping rule and k, a large number of data sets were simulated to evaluate the RMSE, bias and variance of the estimators. Given the log-likelihood ℓ_{M} for $\mathsf{M} \in \{\text{std}, \text{ex}, \text{fc}, \text{pc}\}$, detailed in equations (2.1), (2.2), (3.1) and (3.2), profile likelihood confidence intervals for a return-level are studied in terms of their coverage and width.

One major issue with simulating data sets with stopping rules is parameter identifiability. For observation i, the stopping decision of the variable-threshold rule is based on the parameter MLEs using observations $1, \ldots, i - 1$. However, with $N \leq 2$ observations contributing to a likelihood the GEV parameters are strictly not identifiable, and for larger but low values of N the parameters are still not practically estimable. As discussed in §2.2, in practice there is typically additional information which is incorporated into decisions, and our analysis should allow for this also. Such historical information is treated as *fixed* and introduces a fixed extra penalty term, $P_{hist}(\theta)$, into the log-likelihood; in a Bayesian analysis it would constitute prior information about the parameter vector. As our simulation studies are conducted without such evidence we treat the first n_0 simulated values as providing historical information on θ; we call $\tilde{x} := (x_1, \ldots, x_{n_0})$ the *historical data.* Thus each simulated data set has the additive penalty contribution to the log-likelihood:

$$P_{hist}(\theta) = \sum_{j=1}^{n_0} \ell(\theta; x_j), \tag{4.1}$$

a contribution that does not depend on the stopping rule since we imagine that these data were available *before* decisions to stop and analyse the data were being made.

We fix the historical data, $\tilde{x}$, using an even spread of values:

$$x_j = G^{-1}(j/(n_0+1); \theta) \qquad \text{for} \quad j = 1, ..., n_0 \tag{4.2}$$

where G is the distribution function of the data-generating GEV distribution. In addition to providing a natural spread of values and stabilising the likelihood, for the fixed-threshold rule, provided c_k is greater than the $1/(n_0+1)$ return-level, no historical value exceeds the stopping threshold. The stopping threshold, $\hat{x}_k(\boldsymbol{x}_{1:i-1})$ is now, implicitly, also a function of $\tilde{x}$. We take $n_0 = 10$, the smallest value that gave reliable numerical estimates for $\hat{x}_k(\boldsymbol{x}_{1:i})$ with $i \geq n_0$, across the set of different true values for θ that were used in the simulation study.

4.2 Fixed threshold stopping rule

In the appendix §B we describe in detail the behaviour of the shape parameter estimates in our simulations. In particular we found that the standard shape parameter estimator has both large positive bias and large variance. The formulae in §3.1 show that return level estimates are exponential in the shape parameter. For high return periods moderately large ξ estimates can lead to unrealistically high return-level estimates which exert unwarranted influence on statistics based on empirical averages, such as estimated bias. Hence, we use trimmed averages here.

Figure 3 shows the relative bias and RMSE of the 200 year return-level estimators when sampling using the fixed-threshold stopping rule from the GEV distribution with $\xi = 0.2$. Similar sets of plots for the 50 and 1000 year return-level estimator and $\xi = -0.2$ can be found in the supplementary material. The main driver of RRMSE in all cases is found to be the variance of the estimators, so changes in bias are not too important in this regard. Overall, the return-level estimator which results in the lowest RRMSE most consistently is $\hat{x}_y^{pc}$, mostly due to the low variance of these estimates whereas $\hat{x}_y^{fc}$ has the lowest bias. Both conditioning estimators, $\hat{x}_y^{pc}$ and $\hat{x}_y^{fc}$, improve upon the $\hat{x}_y^{std}$ especially when we are estimating very high return-levels (*i.e.*, for larger y) and/or the underlying distribution is heavy tailed. Although $\hat{x}_{200}^{ex}$ has somewhat similar properties to $\hat{x}_{200}^{pc}$ for $\xi = 0.2$ it has larger RRMSE for $\xi = -0.2$. The fitted distribution using ℓ_{ex} typically has a lighter tail and can even have an upper end point which is less than the excluded observation.

The coverage for all likelihoods gets closer to the correct value (here 95%) as k increases for any return period, y. For ℓ_{std} we have overcoverage and the widest confidence intervals on average and using ℓ_{fc} we have good coverage, particularly when the distribution is heavy tailed. For the other likelihoods there is mostly undercoverage (coverage ranging from 80-95%) due to upper bounds being too low. The exclusion of upper tail information results in relatively narrow confidence intervals from ℓ_{ex}. In contrast, ℓ_{fc} produces a higher upper confidence limit and hence a wider confidence interval than ℓ_{pc} and ℓ_{ex} because the likelihood essentially neglects

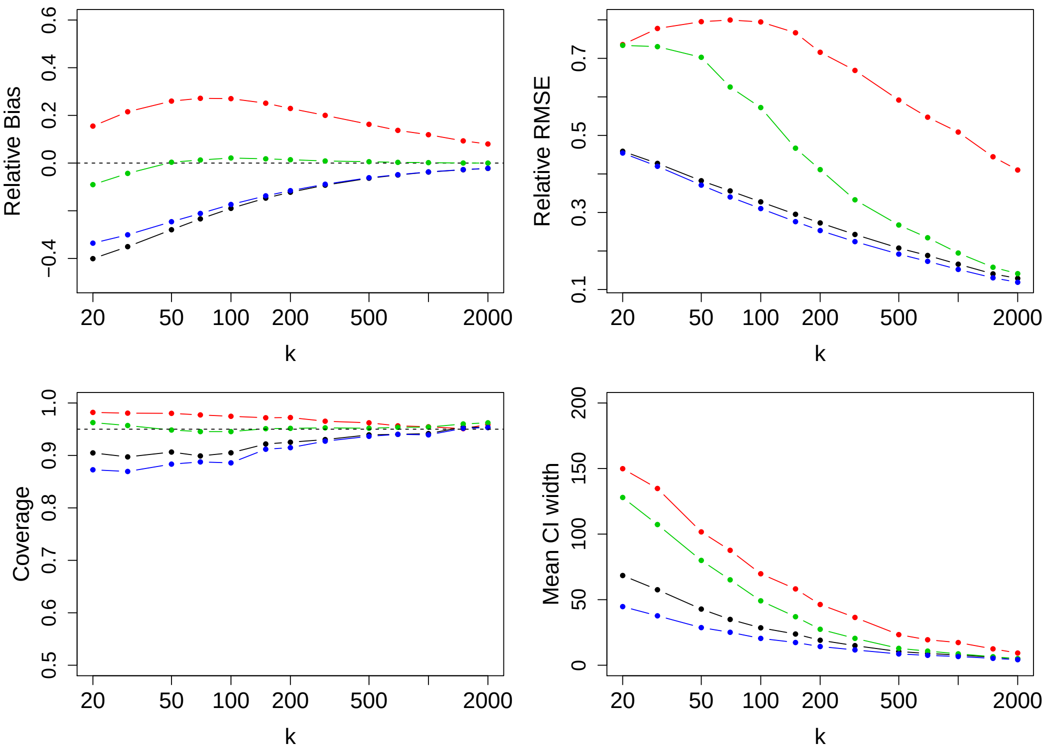

Figure 3: *200 year return-level estimates when sampling from the GEV distribution with* $(\mu, \sigma, \xi) = (0, 1, 0.2)$ *using the fixed-threshold stopping rule over a range of thresholds. From left to right. Top: relative bias and relative RMSE. Bottom: coverage and average CI width. Colour scheme is the same as in Figure 2. Based on* 10^5 *replicated samples with the historical data created using approach* (4.2)*. Coverage is based on 5000 replicated samples.*

the distribution of N, which contains some information about the upper tail of the distribution. The confidence intervals produced using ℓ_{fc} vary greatly in width across our simulations with a larger median width than those using ℓ_{std}.

4.3 Variable threshold stopping rule

Within the samples simulated we find the stopping thresholds, $\hat{x}_k(x_{1:m})$, over $m < N$ are generally less than the true k-year return level, x_k. As a result the samples are both smaller in size and consist of smaller values than when using the fixed-threshold stopping rule. So return level estimates calculated using ℓ_{std} have a small positive bias and those calculated using the other three likelihoods have a larger negative bias than observed for the fixed-threshold stopping rule.

The properties of the 200 year return-level estimators for $\xi = 0.2$ are shown in Figure 4, the 50 and 1000 year return-levels and $\xi = -0.2$ are considered in the supplementary material (Figures S.12-S.17). For $\xi = 0.2$ the conditioning methods provide the best return-level estimators in terms of RMSE despite the estimators having a larger squared bias than $\hat{x}_y^{std}$. The reason for this is that for heavy tailed distributions the variances of return-level estimators are generally larger than the bias. However, for lighter tailed distributions the bias plays a larger role as the relative variances of the different estimators are much closer together. As a result $\hat{x}_{200}^{pc}$ and $\hat{x}_{200}^{fc}$ perform marginally worse than $\hat{x}_{200}^{std}$ in terms of RRMSE when the distribution has a light tail.

The coverage for ℓ_{std} is high (96-98%), decreasing only slightly as k increases but it has the widest confidence intervals generally. Using either ℓ_{pc} or ℓ_{ex} leads to undercoverage, as k increases

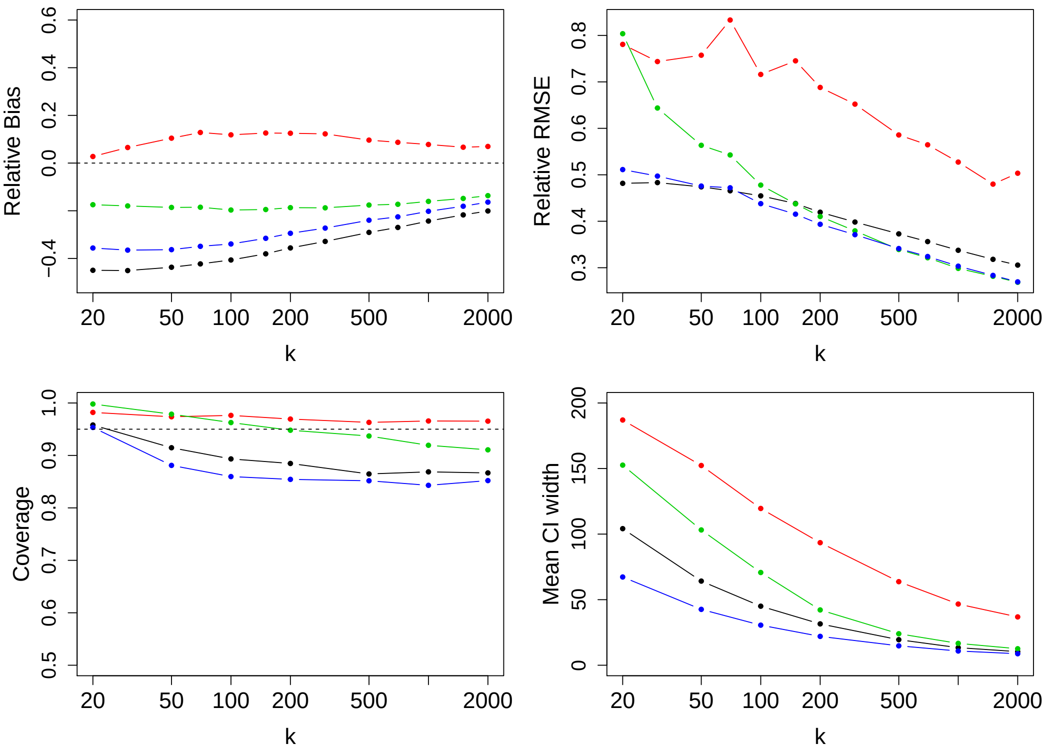

Figure 4: *200 year return-level estimates when sampling from the GEV distribution with* $(\mu, \sigma, \xi) = (0, 1, 0.2)$ *using the variable-threshold stopping rule over a range of* k*. See Figure 3 for associated detail. Based on* 10000 *replicated samples with the historical data created using approach* (4.2)*. Coverage is based on 3000 replicated samples.*

ranging from approximately 95% to 83-87% for ℓ_{ex} and from 93% to 78-85% for ℓ_{pc} with coverage higher when the distribution is heavy tailed. The coverage for ℓ_{fc} also reduces with increasing k from 99-100% for $k = 20$ to approximately 90% for larger k. On average the confidence intervals using ℓ_{fc} are narrower than using ℓ_{std} but generally wider than those using ℓ_{ex} or ℓ_{pc}.

Overall, ℓ_{fc} provides the 'best' results when using the variable-threshold stopping rule. The RRMSE of $\hat{x}^{fc}_{200}$ is generally lower than that of $\hat{x}^{std}_{200}$, coverage is above 90% and the confidence intervals are narrower on average than those using the ℓ_{std}. Although ℓ_{pc} provides estimators with a lower RRMSE than ℓ_{fc}, particularly when the distribution is heavy tailed, it has more severe undercoverage.

4.4 Use in Practice

In practice, for the analysis of data that we believe has been obtained by the flood management agencies using a the fixed-threshold rule we must set a threshold, c, and if they use a variable-threshold rule we must set a return period, k, neither of which may be known. This is important since the behaviour of the estimators can vary depending on the return period, k, associated with the stopping threshold (as we have seen in §4.2 and 4.3). For the fixed-threshold rule, c should lie between $\max_{i<n} x_i$ and x_n. For the variable-threshold rule k should be such that $x_i \leq \hat{x}_k(x_{1:i-1})$ for all $i < n$, but $x_n > \hat{x}_k(x_{1:n-1})$. To use the simulation study results to understand the properties of the estimators it is useful to narrow down a range of feasible k. For the variable rule, a range of possible k can be determined from the data. However, for the fixed-threshold stopping rule k is unknown. Nevertheless, we are likely to have some idea of the range of k which corresponds to c, *i.e.*, we have a prior belief for k.

The 'historical data' also needs to be determined, maybe incorporating prior knowledge in some way. The simplest approach is to start using the stopping rule after the first n_0 observations of the data set and use these n_0 values as the historical data. The choice of n_0 only affects the point estimates and confidence intervals using ℓ_{fc}. However n_0 and the historical data itself can have a large impact on the properties of the estimators, particularly when the sample size is small. In the simulation study, out of necessity, we have restricted ourselves to a particular fixed historical sample, so for low k the properties of the estimators will differ slightly in practice.

5 Case Study - Lune at Caton

We now consider the analysis of the 48 annual maximum river flow observations from the Lune at Caton introduced in §1. Figure 1, right panel, shows the inference for the 200 year return-level of the data, at yearly intervals as new data are observed, with the analysis not accounting for any stopping rule. We now estimate this return-level using the four inference methods (standard, exclude, and our full- and partial-conditional) for both fixed- and variable-threshold stopping rules for a range of levels (c and k respectively), where we drop the subscript of c as the return period of the stopping threshold is unknown. The following discussion assumes that the sampling procedure is well approximated by these respective stopping rules for the selected c and k. In all cases we take the historical data to be the first $n_0 = 10$ observations as in practice no estimates of long period return levels would be attempted from smaller samples. We also consider the implications if a trend in the annual maxima is also simultaneously estimated.

5.1 Fixed-threshold stopping rule

First we discuss the inference using the fixed-threshold stopping rule with $c = 1568m^3/s$, where, for illustration purposes, c is taken to be the mid-point between the 1995 and 2015 levels and the realised value of N is 38, *i.e.*, we stop after 2015. Figure 1, right panel, to the right of the vertical dotted line, shows the estimates and the associated 95% confidence intervals for the four inference methods. The estimates $\hat{x}_{200}^{std}$ and $\hat{x}_{200}^{ex}$ are identical to the estimates in the right panel of the figure for years 2015 and 2014 respectively. Both $\hat{x}_{200}^{fc}$ and $\hat{x}_{200}^{pc}$ (evaluated at 2015) are only slightly larger than the $\hat{x}_{200}^{std}$ estimates for the years before 2015 and $\hat{x}_{200}^{ex}$. From §4.2 we know that, when employing the fixed-threshold stopping rule, $\hat{x}_{200}^{std}$ is positively biased, $\hat{x}_{200}^{fc}$ is close to being unbiased and both $\hat{x}_{200}^{ex}$ and $\hat{x}_{200}^{pc}$ have some negative bias, therefore it is reassuring to see that for 2015, $\hat{x}_{200}^{std} >> \hat{x}_{200}^{fc} > \hat{x}_{200}^{pc} > \hat{x}_{200}^{ex}$.

The confidence interval for 2015 using ℓ_{std} is wider than the intervals of the previous 15 years, especially the 2014 interval (*i.e.*, using ℓ_{ex}), and both the lower and upper bounds are much larger. In this case study, the confidence interval of x_{200} using ℓ_{pc} is similar but slightly narrower than when using ℓ_{ex}. However using ℓ_{fc} the interval is wider (since the upper bound increases) than if we just ignored the 2015 event (using ℓ_{ex}) so we are capturing some of the increased uncertainty in the heaviness of the tail that this event has caused. Nevertheless, the upper confidence bound of x_{200} obtained using ℓ_{fc} is lower than that obtained using ℓ_{std}.

The behaviour of the confidence intervals of these methods appears to be in line with our coverage and width results in §4.2. Indeed, here the shape parameter estimates, $(\hat{\xi}_{std}, \hat{\xi}_{ex}, \hat{\xi}_{fc}, \hat{\xi}_{pc})$, are $(0.04, -0.07, -0.04, -0.05)$ so we expect coverage to be between the coverage values found in the simulation study for $\xi = 0.2$ and $\xi = -0.2$. In the study we found that using ℓ_{std} with the fixed-threshold stopping rule leads to overcoverage (95-98% for $\xi = 0.2$, 97.5-99.5% for $\xi = -0.2$) and the upper bound of the confidence interval found using ℓ_{std} is lower than x_{200} only 1-2% of the time, so it is likely that for the Lune data the upper bound of the confidence interval using

ℓ_{std} is too high. This is further emphasised for the Lune estimates by the upper bound for 2015 exceeding the associated values for the previous 30 years (Figure 1). In §4.2 we found that $\hat{x}_{200}^{ex}$ and $\hat{x}_{200}^{pc}$ exhibited narrow confidence intervals which together with their negative bias led to undercoverage, with the upper bounds being too low, especially when $\xi = -0.2$ and k is low. For our chosen $c = 1568$ we can obtain estimates of the corresponding return period, k, of c; in particular $\hat{k}_{std} = 90$ and $\hat{k}_{ex} = 550$ and we expect k to lie between these two values. Thus, using the simulation study results, we expect that the coverage of the ℓ_{pc} and ℓ_{ex} confidence intervals to lie between 85 and 95%. However the lower bounds of these confidence intervals were found to be less than x_{200} for almost 100% of simulated samples so it is highly likely that the true 200-year return level for the Lune data is above the lower bounds given by the ℓ_{pc} and ℓ_{ex} confidence intervals. For ℓ_{fc} and $90 < k < 550$, the coverage is 94-95% with the percentage of upper bounds too low being 3-6% suggesting that with the Lune data the upper bound of the ℓ_{fc} confidence interval is likely to be higher than the true 200-year return level, x_{200}.

The above discussion assumed that c was known. In some cases this may be true as c could represent a known physical limit linked to flooding. This is not the case for the Lune at Caton, with our value chosen subjectively for illustrative purposes although it could be argued that lower c values in this range would be more reasonable since the 1995 river flow observation was considered high as it led to flooding. To assess the impact of c we consider a range of values for c between the 1995 and 2015 observations, with the inference for the four methods presented in Figure 5, left panel (when $c = 1568$ the estimates are those shown in Figure 1 right panel). Now, $\hat{x}_{200}^{std}$ and $\hat{x}_{200}^{ex}$ and the corresponding confidence intervals are invariant to c but as c increases $\hat{x}_{200}^{fc}$ and $\hat{x}_{200}^{pc}$ both decrease. As noted earlier, $\hat{x}_{200}^{fc} > \hat{x}_{200}^{pc}$ but they become closer as c tends to the 2015 event level because the information that ℓ_{fc} discards, *i.e.*, the probability of the event that c was not exceeded on the first $n-1$ observations, becomes less informative. The confidence intervals using the conditioning likelihoods notably narrow with increasing c; the lower bounds slightly decrease but the largest reduction is in the upper bounds. For lower c values the ℓ_{fc} intervals are wider than for ℓ_{std}, in contrast for the largest possible c values the interval is very narrow (a reduction in size of factor 14 over the range of c possible). For ℓ_{pc} the upper bounds are smaller than those using the ℓ_{std} for all values of c and are slightly larger than those for ℓ_{ex} for low c. However, for large c the upper bounds of both conditioning confidence intervals are much lower than that using ℓ_{ex} since the information that c was exceeded on this observation becomes more informative about the tail of the distribution as c approaches the 2015 observation. Thus if we stop after the first minor exceedance of c we can be reasonably sure the tail is short. This is an unexpected but helpful finding. Further investigation into the confidence intervals can be found in Barlow (2020).

5.2 Variable-threshold stopping rule

Now we consider the variable-threshold stopping rule and first determine a range of k from the data. In the Lune data the maximum river level in 2015 corresponds to $\hat{k} = 2561$ given the data up to 2015 and to $\hat{k} = 188$ using all the data. However, the river level in 1995 corresponds to $\hat{k} = \infty$ (*i.e.*, it is larger than the point estimate of the upper end point of the GEV fitted to the data up to 1995) so the variable threshold rule as given in (1.4) cannot have been applied for any $k < \infty$. Furthermore, the river level in 1980 corresponds to $\hat{k} = 111$. If the variable-threshold stopping rule had motivated a request for an analysis of the data up to and including 2015, the request must have been triggered by the second such exceedance. In our analysis we explore values of k between 200 and 2500 and simply amend ℓ_{fc} slightly by replacing the ℓ_{fc} contribution of the 1995 observation ($i = 28$), $g(x_{28}; \theta)/G(\hat{x}_k^{std}(\boldsymbol{x}_{1:27}); \theta)$, by $g(x_{28}; \theta)$.

Figure 5, right panel, shows the same inferences as the left panel, but for the variable-threshold over a range of return periods $k \in [200, 2500]$. Given the rarity of all events in this range we

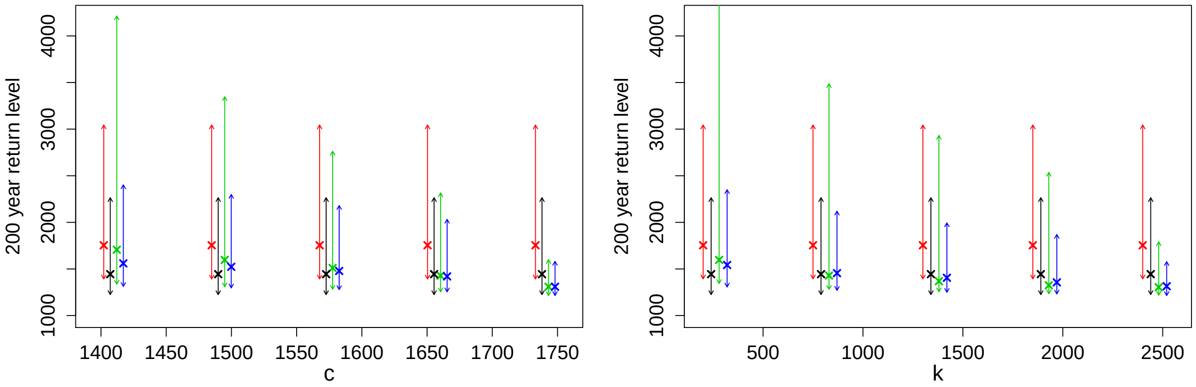

Figure 5: *200-year return-level estimates based on all the data up to and including 2015 for the Lune at Caton with 95% profile likelihood-based confidence intervals: left with the fixed-threshold stopping rule over a range of c and right with the variable-threshold stopping rule over a range of k: standard likelihood (red), excluding the final observation (black), full conditioning (green) and partial conditioning (blue). Each group of 4 estimates applies for the same c/k as for the standard estimate in each group and have been horizontally shifted for clarity.*

would expect a 'true' k to be towards the lower end of this range. The estimates $\hat{x}_{200}^{std}$ and $\hat{x}_{200}^{ex}$ and the corresponding confidence intervals are invariant to k (and independent of the stopping rule used) but as k increases $\hat{x}_{200}^{fc}$ and $\hat{x}_{200}^{pc}$ both decrease. For small k, $\hat{x}_{200}^{fc} > \hat{x}_{200}^{pc}$, as we would expect from our bias results in the simulation study. However, the inequality reverses for large k perhaps as a result of there being more than one exceedance of the threshold. This is hinted at by the bias results and also since if one omits the 1995 observation from the data set then $\hat{x}_{200}^{fc} > \hat{x}_{200}^{pc}$ for all k. More investigation into the estimators when there are multiple exceedances would be useful.

The intervals using the conditioning likelihoods and variable-threshold stopping rule behave similarly to those using the fixed-threshold stopping rule. Again the ℓ_{fc} intervals are highly influenced by the 'extremeness' of the stopping threshold. With the lowest possible k for this data set (ignoring the 1995 exceedance) the ℓ_{fc} interval is more than double the width of the confidence interval using ℓ_{std} whereas for a large k value it is less than half the width. The ℓ_{pc} confidence intervals also reduce in width with increasing k but not as dramatically.

5.3 Non-stationarity

The implications of using stopping rules on the estimation of trends in extreme levels is also a concern, as stopping with the final observation being large is likely to have a similar biasing effect as found in §2 and §4 for return-levels. This is particularly important given the interest in whether trends in extreme values differ from trends in mean levels (Eastoe and Tawn, 2009; Hannaford and Marsh, 2008). In Figure 6 we illustrate the analysis of the Lune data with a GEV distribution including a linear trend $\mu_t = \alpha_0 + \beta t$, showing both the resulting estimates of the 200 year return-level for 2015, *i.e.*, the estimates of the 0.995 quantile of the annual maximum in 2015, and the associated trend estimate $\hat{\beta}$ using progressively more data over time. With few data used the trend is estimated to be unrealistically large, with huge uncertainty, and this results in very different point estimates of return-levels relative to the analysis with no trend. As more data are observed we can see that the trend estimates generally decrease, with reduced uncertainty, with positive jumps in $\hat{\beta}$ estimates after the large 1995 and 2015 events. Although the 2015 river flow is more extreme than that of 1995 its impact on $\hat{\beta}$ is much less. Furthermore,

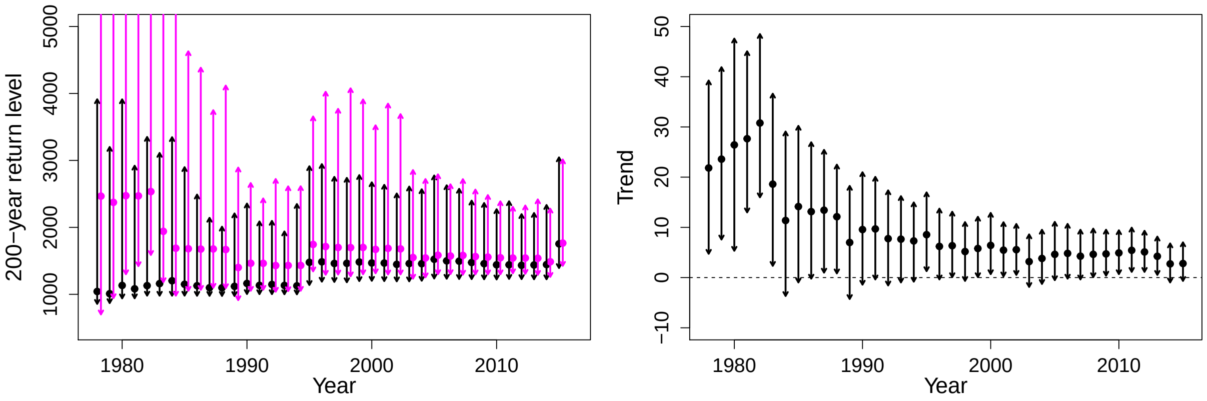

Figure 6: *Fitting a GEV to all the data up to and including the current year for the Lune at Caton. Left: 200yr return-level estimates for* 2015 *using progressively more data over the years with and without a trend in the location parameter (pink and black respectively). Each group of 2 estimates applies for the same year and have been horizontally shifted for clarity. Right: Slope parameter, $\hat{\beta}$, and it's 95% confidence interval.*

we see that $\hat{\beta}$ is not larger than $\beta = 0$ at the 2.5% significance level. Thus here the effect of including the estimated trend is small on the 200 year return-level estimate and the stopping rule seems to have almost no effect on the trend estimate.

6 Discussion

In this paper and the associated supplementary material we have shown that return-level estimators based on the standard likelihood are positively biased when sampling from the GEV or GP distributions using certain stopping rules. The extent of the stopping bias is lower for lighter tailed distributions and when estimating low return-levels. We have proposed conditioning upon the stopping threshold in the likelihood. In most cases we have found that conditioning on the final observation exceeding the stopping threshold (partial conditioning) results in return-level estimates with the lowest RMSE despite the estimator being negatively biased.

A balance must be struck between low RMSE and good coverage, however. Partial conditioning results in undercoverage despite the low RMSE of $\hat{x}_y^{pc}$. The full-conditional likelihood, which also conditions on the non-exceedance of all previous observations, gives the closest to 95% coverage and though the intervals are wide, they are typically narrower than the confidence intervals obtained from the standard likelihood. The interval widths using the full and partial conditional likelihoods are smaller the closer the stopping threshold is to the final observation as the occurrence of the final exceedance becomes more informative on the tail of the distribution (see §5.1).

Overall, the conditioning estimators presented here outperform the standard estimator when the decision to analyse data at a particular time was triggered by what was perceived to be a large observation. For the fixed-threshold stopping rule, partial conditioning has the best combination of RMSE and coverage for a range of ξ with moderate k and particularly when the distribution is heavy tailed, as is the case for most UK rivers (Institute of Hydrology, 1999). For the variable-threshold stopping rule, full conditioning provides the best balance of coverage and low RMSE. To apply the conditioning estimators in practice if the rule of the flood management agency is unknown the statistician needs to choose a suitable stopping threshold, c, for the fixed-threshold stopping rule and a suitable stopping 'period', k, for the variable-threshold stopping rule if the values are unknown. A range of c and k can be considered provided that the observed data are below the resulting stopping threshold(s) up to the final observation.

The decision to analyse data will likely be based on a confluence of many factors. Our work attempts to simplify the true decision making procedure by using stopping rules based on the occurrence of a *single* large observation exceeding some threshold. An analysis may instead be prompted by a prolonged period of quite large (but not necessarily 'extreme') observations or the observation of large values at many locations simultaneously, requiring more complex multivariate analysis since the observations at nearby locations will be dependent in some way (Keef et al., 2009; Asadi et al., 2015).

In practice if the stopping rule is unknown and the analysis is triggered by a large event, we suggest using the full conditional return-level estimator. However if k is thought to be less than 50, or the full-conditional estimate and/or confidence interval are clearly too large then partial conditioning should be used instead. We argue that the decision to 'stop' and analyse data would in part be based on both past return-level estimates and thresholds set due to current infrastructure and so the 'true' stopping rule is a mixture of the two rules considered here. Hence the 'true' bias, RMSE and coverage of the estimators can be expected to lie between those which we found under the two stopping rules. It should be noted that this work does not address the question of *when* the data should be analysed, but rather how we can reduce the bias given the use of a particular stopping rule. Nevertheless, if we are at a point in time where a stopping criterion has been met and triggered an analysis, this study can give guidance on the behaviour of return-level estimators calculated at the current time whether based on the full likelihood, partial or full conditioning, or even excluding the most recent, 'triggering' event.

In our theoretical and simulation studies we have not accounted for the possibility of a trend in the data, such as river flows gradually increasing over the years. We saw in §5 that the Lune data has a slight positive trend in the location parameter and fitting such a model at an earlier point in time resulted in a very large positive trend. This could cause problems for the fixed-threshold stopping rule, in particular it might become necessary to change c after a certain number of years. Nonetheless, doing this is probably not too unrealistic since, for example, the height of a flood defence might be increased if there has been evidence of higher flow in recent years. On the other hand the variable-threshold stopping rule is more robust to data with an underlying trend as it is directly a function of the observed data.

Acknowledgements

Barlow gratefully acknowledges funding of the EPSRC funded STOR-i centre for doctoral training (grant number EP/L015692/1). This research was also financially supported by JBA. The authors also thank NRFA for the Lune gauge data.

References

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *Annals of Applied Statistics*, 9(4):2023–2050.

Barlow, A. M. (2020). Extreme Value Problems in Hydrology. *PhD thesis (in preparation), Lancaster University.*

Barndorff-Nielsen, O. E. and Cox, D. R. (1984). The effect of sampling rules on likelihood statistics. *International Statistical Review*, 52(3):309–326.

Barry, J. and Coggan, R. (2010). The visual fast count method: Critical examination and development for underwater video sampling. *Aquatic Biology*, 11(2):101–112.

Bauer, P., Koenig, F., Brannath, W., and Posch, M. (2010). Selection and bias—two hostile brothers. *Statistics in Medicine*, 29(1):1–13.

Bowden, J. and Glimm, E. (2008). Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal*, 50(4):515–527.

Carreras, M. and Brannath, W. (2013). Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine*, 32(10):1677–1690.

CEH, N. (2018). National River Flow Archive.

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

Cox, D. R. (1952). A note on the sequential estimation of means. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(03):447.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 52:393–442.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):25–45.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Efron, B. (1990). More efficient bootstrap computations. *Journal of the American Statistical Association*, 85:79–89.

Eggen, T. J. H. M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika*, 65(3):337–362.

Hannaford, J. and Marsh, T. J. (2008). High-flow and flood trends in a network of undisturbed catchments in the UK. *International Journal of Climatology*, 28(10):1325–1338.

Institute of Hydrology (1999). *Flood Estimation Handbook*, volume 3.

Keef, C., Svensson, C., and Tawn, J. A. (2009). Spatial dependence in extreme river flows and precipitation for Great Britain. *Journal of Hydrology*, 378(3-4):240–252.

Kenward, M. G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13(3):236–247.

Molenberghs, G., Kenward, M. G., Aerts, M., Verbeke, G., Tsiatis, A. A., Davidian, M., and Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: Sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, 23(1):11–41.

Pawitan, Y. (2013). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. OUP Oxford.

Stallard, N. and Todd, S. (2005). Point estimates and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference*, 135(2):402 – 419.

Todd, S. (2007). A 25-year review of sequential methodology in clinical studies. *Statistics in Medicine*, 26(2):237–252.

Todd, S., Whitehead, J., and Facey, K. (1996). Point and interval estimation following a sequential test. *Biometrika*, 83(2):453–461.

Wald, A. (2004). *Sequential Analysis (reprint).* Dover Phoenix Editions. Dover Publications.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73(3):573–581.

Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medicine*, 18(17-18):2271–2286.

Appendix

A Proof of results from §2

A.1 Proof of Proposition 1

For simplicity we denote c_k by c. Sampling from some general distribution with the first stopping rule, we have:

$$\mathbb{E}\left[\frac{1}{N}\right] = \sum_{n=1}^{\infty} \bar{F}(c)\frac{F(c)^{n-1}}{n} = -\frac{\bar{F}(c)}{F(c)}\log(\bar{F}(c))\,, \tag{A.1}$$

where $F(x)$ and $\bar{F}(x) = 1 - F(x)$ are the CDF and survival function of the distribution of $X_i,\ i = 1, \ldots, n$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\overline{X}_N\right] &= \mathbb{E}\left[\frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} X_i | N = n\right]\right] \\
&= \mathbb{E}\left[\frac{1}{N}\left((N-1)\mathbb{E}\left[X|X \le c\right] + \mathbb{E}\left[X|X > c\right]\right)\right] \\
&= \mathbb{E}\left[X|X \le c\right] + \mathbb{E}\left[\frac{1}{N}\left(\mathbb{E}\left[X|X > c\right] - \mathbb{E}\left[X|X \le c\right]\right)\right] \\
&= \mathbb{E}\left[X|X \le c\right] + \mathbb{E}\left[\frac{1}{N}\right]\left(\mathbb{E}\left[X|X > c\right] - \mathbb{E}\left[X|X \le c\right]\right).
\end{aligned} \tag{A.2}$$

Specifically for sampling from the exponential distribution:

$$\mathbb{E}\left[\frac{1}{N}\right] = \frac{\beta c}{e^{\beta c} - 1}.$$

By the memoryless property of the exponential distribution:

$$\mathbb{E}\left[X|X > c\right] = \mathbb{E}\left[X\right] + c = \frac{1}{\beta} + c\,,$$

and rearranging $\mathbb{E}\left[X\right] = F(c)\mathbb{E}\left[X|X \le c\right] + \bar{F}(c)\mathbb{E}\left[X|X > c\right]$ gives

$$\mathbb{E}\left[X|X \le c\right] = \frac{1}{F(c)}\left[\frac{1}{\beta} - \bar{F}(c)\left(c + \frac{1}{\beta}\right)\right] = \frac{1}{\beta} - c\frac{\bar{F}(c)}{F(c)}\,. \tag{A.3}$$

Therefore, for the exponential distribution,

$$\mathbb{E}\left[\overline{X}_N\right] = \frac{1}{\beta} + \frac{c}{e^{\beta c} - 1}\left(\frac{\beta c}{1 - e^{-\beta c}} - 1\right). \tag{A.4}$$

For the standard estimator based on the full sample we have $1/\hat{\beta}_{std} = \overline{X}_N$, the sample mean. The first part of Proposition 1 then follows from (A.4).

If the final data point is excluded from the sample then all included samples are from the distribution truncated at c, so, from (A.3),

$$\mathbb{E}\left[\frac{1}{\hat{\beta}_{ex}}\middle| N > 1\right] = \mathbb{E}\left[X | X \leq c\right] = \frac{1}{\beta} - \frac{c}{e^{\beta c} - 1},$$

leading to the expression in the second part of Proposition 1.

A.2 Proof of Theorem 1

We start by defining the following key quantities for each $k \geq 1$,

$$S_k := (n_0 + k)\overline{X}_k = n_0\overline{X}_0 + \sum_{j=1}^{k} X_j, \tag{A.5}$$

$$V_k := \frac{X_k}{S_k}. \tag{A.6}$$

Marginally $S_k \sim \mathsf{Gamma}((n_0 + k)\alpha, \beta)$ and $V_k \sim \mathsf{Beta}(\alpha, (n_0 + k)\alpha)$; we denote their marginal densities as:

$$f_{S_k}(s_k) \propto s_k^{(n_0+k-1)\alpha-1} e^{-\beta s_k}$$
$$f_{V_k}(v_k) \propto v_k^{\alpha-1}(1 - v_k)^{(n_0+k-1)\alpha-1}.$$

The stopping time, N, is n if $X_n > \gamma\overline{X}_{n-1}$ and $X_i < \gamma\overline{X}_{i-1}$ for $1 \leq i < n$. However, $X_n > \gamma\overline{X}_{n-1}$

$$\begin{aligned}
\Leftrightarrow X_n &> \frac{\gamma}{n + n_0 - 1}(S_n - X_n) \\
\Leftrightarrow \left(1 + \frac{\gamma}{n + n_0 - 1}\right) X_n &> \gamma \frac{1}{n + n_0 - 1} S_n \\
\Leftrightarrow \left(1 + \frac{\gamma}{n + n_0 - 1}\right) V_n &> \frac{\gamma}{n + n_0 - 1} \\
\Leftrightarrow \qquad V_n &> \frac{\gamma}{n + \gamma + n_0 - 1}.
\end{aligned}$$

So the stopping rule can be written purely as function of the Vs. Explicitly, we stop at time n if $V_n > \frac{\gamma}{n+\gamma+n_0-1}$ and $V_i < \frac{\gamma}{i+\gamma+n_0-1}$ for $1 \leq i < n$.

We define the statement $\mathcal{A}_n :=$ "$V_1, \ldots, V_n, S_n$ are mutually independent". Below, we will show by induction that $\mathcal{A}_n$ holds for all $n \geq 1$. Thus $\overline{X}_n \perp\!\!\!\perp V_i \quad \forall i \leq n$; the distribution of $\overline{X}_n$ is independent of whether or not the stopping rule has been triggered. Therefore, $\overline{X}_N$ conditioned on $N = n$ is equivalent to the mean of n i.i.d. $\mathsf{Gamma}(\alpha, \beta)$ random variables, as stated in the theorem.

$\underline{\mathcal{A}_{n-1} \Rightarrow \mathcal{A}_n}$: If $\mathcal{A}_{n-1}$ holds then the joint pdf of $V_1, \ldots, V_{n-1}, S_{n-1}$ can be factorised:

$$f_{n-1}(v_1, \ldots, v_{n-1}, s_{n-1}) = f_{S_{n-1}}(s_{n-1}) \prod_{i=1}^{n-1} f_{V_i}(v_i)\,,$$

Consider the change of variables $(V_1, \ldots, V_{n-1}, S_{n-1}, X_n) \to (V_1, \ldots, V_{n-1}, V_n, S_n)$, where $X_n = S_n V_n$ and $S_{n-1} = S_n(1 - V_n)$. The Jacobian for this transformation is:

$$|J| = \left| \frac{\partial(v_{1:n-1}, s_{n-1}, x_n)}{\partial(v_{1:n}, s_n)} \right| = \begin{vmatrix} I_{n-1} & 0 \\ 0 & A \end{vmatrix} = s_n\,.$$

where I_{n-1} is the $(n-1) \times (n-1)$ identity matrix and

$$A = \frac{\partial(s_{n-1}, x_n)}{\partial(v_n, s_n)} = \begin{bmatrix} -s_n & 1 - v_n \\ s_n & v_n \end{bmatrix}.$$

So, since S_{n-1} and $V_1, \ldots, V_{n-1}$ are independent of X_n,

$$\begin{aligned} f_n(v_{1:n}, s_n) &= f_{n-1}(v_{1:n-1}, s_{n-1}(s_n, v_n)) f_X(x(s_n, v_n))|J| \\ &\propto \left(\prod_{i=1}^{n-1} f_{V_i}(v_i) \right) (s_n(1-v_n))^{(n_0+n-1)\alpha-1} e^{-\beta s_n(1-v_n)} (s_n v_n)^{\alpha-1} e^{-\beta s_n v_n} s_n \\ &= \left(\prod_{i=1}^{n-1} f_{V_i}(v_i) \right) s_n^{(n+n_0)\alpha-1} e^{-\beta s_n} v_n^{\alpha-1} (1-v_n)^{(n_0+n-1)\alpha-1} \\ &\propto \prod_{i=1}^{n} f_{V_i}(v_i) f_{S_n}(s_n)\,. \end{aligned}$$

So $\mathcal{A}_n$ holds.

<u>$\mathcal{A}_1$ holds</u>: We must show that V_1 and S_1 are independent. We do this by using the change of variables $(\overline{X}_0, X_1) \to (V_1, S_1)$ to show that the joint pdf of V_1 and S_1 factorises.

We have

$$f_{\overline{X}_0, X_1}(\bar{x}_0, x_1) \propto \bar{x}_0^{n_0\alpha-1} e^{-n_0\beta\bar{x}_0} x_1^{\alpha-1} e^{-\beta x_1}$$

and $X_1 = S_1 V_1$ and $\overline{X}_0 = \frac{1}{n_0} S_1(1 - V_1)$. So Jacobian for the transformation is:

$$\begin{aligned} \left| \frac{\partial(\overline{X}_0, X_1)}{\partial(V_1, S_1)} \right| &= \begin{vmatrix} -\frac{s_1}{n_0} & s_1 \\ \frac{1}{n_0}(1-v_1) & v_1 \end{vmatrix} \\ &= \frac{s_1}{n_0} \end{aligned}$$

Thus the joint pdf of V_1, S_1 is:

$$\begin{aligned} f_{V_1,S_1}(v_1, s_1) &\propto s_1(s_1(1-v_1))^{n_0\alpha-1} e^{-n_0\beta(s_1(1-v_1)/n_0} \\ &\qquad (s_1 v_1))^{\alpha-1}\, e^{-\beta s_1 v_1} \\ &= s_1^{(n_0+1)\alpha-1} e^{-\beta s_1} v_1^{\alpha-1} (1-v_1)^{n_0\alpha-1} \\ &\propto f_{S_1}(s_1) f_{V_1}(v_1), \end{aligned}$$

as required. ■

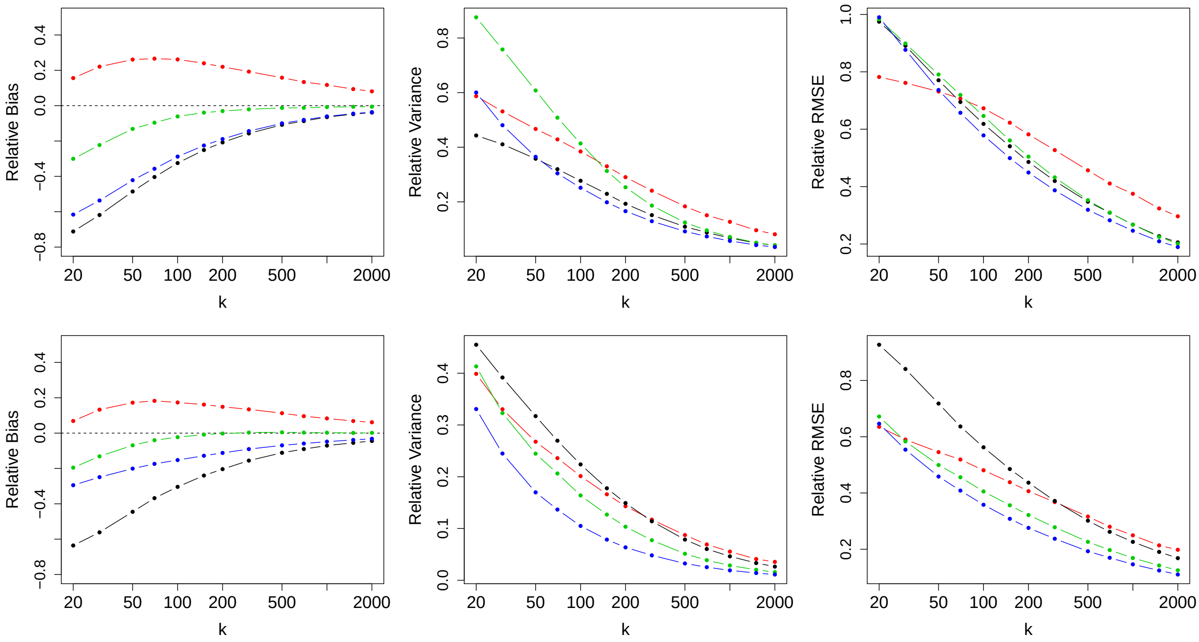

Figure A.1: *Shape parameter estimates when sampling from the GEV distribution with* $(\mu, \sigma) = (0, 1)$ *using the fixed-threshold stopping rule with threshold* c_k *and* $\xi = 0.2$ *(top) and* $\xi = -0.2$ *(bottom) both plotted against* k*. Left: relative bias, centre: relative variance, right: relative RMSE, using: standard likelihood (red), excluding the final observation (black), full conditioning (green) and partial conditioning (blue). Based on* 10^5 *replicated samples with the historical data created using approach* (4.2)*.*

B Properties of the GEV shape parameter

B.1 Fixed-threshold stopping rule

The shape parameter, ξ, is important in determining the tail behaviour. Figure A.1 shows the relative bias, variance and RMSE of each of the estimators when sampling using the fixed-threshold stopping rule for $\xi = 0.2$ and -0.2 (top and bottom rows respectively). Judged by RRMSE, we find that ℓ_{pc} is generally best for moderate to large k, with clear benefits for $\xi = -0.2$; however ℓ_{fc} has generally quite similar RRMSE and low bias. As one would expect the lighter the tail of the distribution, the smaller both the relative variance and, in most cases, the relative bias of the shape parameter estimators resulting in smaller RRMSE. To help understand why these RRMSE results arise we now look at more detail at the performance of the four estimators.

The standard MLE for the shape parameter, $\hat{\xi}_{std}$, is almost always positively biased while $\hat{\xi}_{ex}$ leads to quite large negative bias (with $E(\hat{\xi}_{ex}) < 0.1$ when $\xi = 0.2$ and $k < 50$ (Figure A.1)) since we lose information about the upper tail of the underlying distribution. In particular, the fitted distribution typically has a lighter tail and can even have an upper end point which could be less than the excluded observation. Unlike all other estimators considered, the variance of $\hat{\xi}_{ex}$ is not substantially lower when the tail is lighter and so has quite large RRMSE when $\xi = -0.2$.

The partial conditioning method generally has $\hat{\xi}_{pc}$ lower than the truth however, for moderate k, they consistently have low variance relative to the other methods over a range of ξ. Therefore, partial conditioning provides ξ estimators with the lowest RRMSE for $k > 100$. In contrast, ℓ_{fc} leads to very little bias in ξ estimates for $k > 100$ but the variance can be large, particularly when $\xi = 0.2$ with $k < 100$. This is in agreement with Molenberghs et al. (2014) findings that the full-conditional estimator has poor precision despite it's unbiasedness. However, unlike in Molenberghs et al. (2014), we find that, in our context, full conditioning can improve upon the standard estimator especially when the stopping threshold is high (*i.e.*, for large k).

B.2 Variable-threshold stopping rule

Properties of the shape parameter estimators under the variable stopping rule are shown in the supplementary material, Figure S.11, for $\xi = 0.2$ and -0.2. We find that in the variable threshold setting $\hat{\xi}_{std}$ has very low bias (similarly recall in §2.4 when sampling from the gamma distribution with this stopping rule we found the standard return-level estimator was unbiased) whereas all other ξ estimators are negatively biased, with $\hat{\xi}_{ex}$ having the largest negative bias out of all the estimators for both values of ξ considered. We find that $\hat{\xi}_{std}$ also has the lowest RRMSE of the estimators. Despite $\hat{\xi}_{std}$ performing well under the variable threshold stopping rule, this is not always the case for the ℓ_{std} return-level estimators.