Response to Reviewer 1, R Katz

Thank you for your thorough review. Our responses are in red italics below. Please note that there is considerable overlap with the comments from the other reviewer, and we refer to our separate response to her). We propose to include a new Fig 6 which looks at the CIs as a function of number of bootstrap resamples. We will also include a new paragraph to the Discussion. This is outlined in our reply to Reviewer 2.

GENERAL COMMENTS:

The focus of the manuscript is on efficient use of the bootstrap, a resampling technique, to quantify uncertainty (e.g., in the form of a confidence interval) in estimated extreme statistics such as return levels. Justification is provided for a simplified bootstrap procedure in which the resamples are generated through only drawing from the highest values in the original sample, not the entire sample. This common sense result is consistent with conventional statistical modeling of extremes, with the common assumption that the uncertainty in estimating the rate of exceedance of a high threshold can be ignored (e.g., Chapter 4 in Coles, 2001). Perhaps the present paper serves to place this conventional approach on firmer footing. Nevertheless, there are a number of alternative techniques for uncertainty quantification in extreme value analysis not even mentioned in the manuscript. These alternatives include different implementations of the bootstrap, as well as ones in which no resampling need be performed (e.g., profile likelihood technique; Coles, 2001). At the least, these alternatives should be mentioned.

We agree, and we now mention the weaknesses of non-parametric bootstrapping with reference to Kysely (2008) on page 1, line 13. We also include a brief discussion of the test inversion bootstrap method in the Discussion (page 7, lines 5-13). However, we maintain that this is somewhat beside the point of the article as our main objective has been to investigate how tail statistics can be bootstrapped, if, as the referee says, you must. We do not necessarily argue that non-parametric bootstrapping is the best alternative, and we have made clearer where we think it is appropriate to use (see also our reply to Reviewer 2).

For this reason, I recommend that the manuscript be accepted for publication subject to minor revision.

SPECIFIC COMMENTS: (1) Nonparametric versus parametric bootstrap. A nonparametric bootstrap is used in which the resamples are created by drawing with replacement from the original sample. When fitting extreme value distributions (e.g., the generalized Pareto in Sec. 3.3), it has been suggested that a parametric bootstrap would be preferable for constructing confidence intervals for return levels (i.e., resamples are created by Monte Carlo simulation from the fitted distribution) (Kysely, 2008).

We agree that, especially for small samples, parametric bootstraps are probably capable of better coverage than non-parametric bootstrap techniques. However, we are looking at very large samples, indeed at samples that are so large that we can perform in some cases in-sample estimates of 100-year return values. The paper now acknowledges the limitations of non-parametric bootstraps, and we stress that it should be seen as a study of how to efficiently handle the original data set if, as the reviewers points out, you wish to perform a non-parametric bootstrap. We have included a paragraph in the Discussion (page 7, lines 5-13) where we look at the caveats to using non-parametric bootstraps (see also our reply to Reviewer 2).

(2) Refined bootstrap techniques Bootstrap-based confidence intervals can be too short, especially for return levels with long return periods. Consequently, alternative more involved bootstrap techniques (e.g., the so-called "test inversion" bootstrap) have been proposed to improve the performance of such confidence intervals (Schendel and Thongwichian, 2015).

This is an interesting technique, and Reviewer 2 also refers to a follow-up paper by the same authors. We have included a short paragraph in the Discussion (page 7, lines 5-13) where we outline this alternative method (see also our reply to Reviewer 2).

(3) Alternatives to bootstrap When estimating the parameters of an extreme value distribution by maximum likelihood, an alternative technique for obtaining confidence intervals for return levels is profile likelihood (Coles, 2001). This technique does not require any resampling, but does require repeated fits of the extreme value distribution under parameter constraints. It is competitive with resampling for obtaining confidence intervals of return levels (e.g., Schendel and Thongwichian, 2015).

The profile likelihood technique is a well-known technique, but it falls outside the scope of this paper to investigate it as we focus strictly on efficient methods for non-parametric bootstrapping.

Response to Reviewer 2, S Caires

Thank you for your thorough review. Our responses are in red italics below. Please note that there is considerable overlap with the comments from the other reviewer, and we refer to our separate response to him). We propose to include a new Fig 6 (attached) which looks at the CIs as a function of number of bootstrap resamples. We will also include a new paragraph to the Discussion. See below for details.

General comments:

The authors show that in the non-parametric bootstrap procedure for obtaining confidence intervals of estimates based on the k largest values in a sample, the computations can be carried out in a more computationally efficient way by drawing bootstrap samples from the K0 (K0>k) largest values of the sample rather than from the entire sample. They propose that K0 be fixed at a value leading to a very low probability of drawing fewer than the required k largest entries of the sample and provide the expression of that probability. The article is concise and well-written. The suggested approach appears to be useful for applications such as those considered in examples 1 and 2 (empirical percentile). However, I have doubts about the correctness of the non-parametric bootstrap procedure for obtaining confidence intervals of GPD return value estimates as described in Example 3. I have two major comments that I would like the authors to address or at least consider that they, despite not being covered by the article, should also be taken into consideration when bootstrapping to obtain confidence intervals estimates related to extremes.

Major comments:

1. I was not aware of the idea of the bootstrap being applied to the entire dataset rather than to a sample of cluster peaks as in the computation of confidence intervals of Example 3. In the usual form of the parametric bootstrap one does not return to the entire sample, but considers the (much smaller) sample to which the GPD was fitted. In any case, ensuring that the coverage rates - the percentage of times that a confidence interval really contains the true parameter in (hypothetical) repetitions of the same sampling and estimation process - of bootstrap confidence intervals are sufficiently correct has, in my view, priority over the computational effi- ciency of those intervals. Both Coles and Simiu (2003, J. Engrg. Mech., 129 (11), 1288-1294) and Schendel and Thongwichian (2017, Adv. Water Resour., 99, 53-59, http://dx.doi.org/10.1016/j.advwatres.2016.11.011) consider the shortcomings of bootstrap intervals with respect to coverage, the first paper offering an ad hoc solution and the second suggesting the use of Test Inversion Bootstrap. I wonder if the authors could add information to the article about the coverage rates of their confidence intervals.

The reason we return to the entire sample in Example 3 is that the data set represents independent forecasts (taken at long lead times, as described by Breivik et al, 2013, 2014). We are thus in the situation where we are not limited to a peaks-over-threshold technique but can (and should) resample from the entire sample and then set a threshold (note the difference between a POT and a threshold). It was the magnitude of this data set that motivated us to explore which simplifications can be made in order to speed up the bootstrapping for tail statistics. We have elaborated on this in our revision of Example 3 (see p 5, lines 25-27) to make clearer why it is important to revisit the entire sample. As for the question of whether a non-parametric bootstrapping method will underestimate the width (coverage) of CIs, we agree in general, but note that our examples involve very large data sets. See also our reply to Reviewer 1.

2. The results shown in Figs 3, 4 and 5 are based on M=10,000 bootstrap replications, while those shown in Fig 8 are based on M=1,000. I wonder if the authors could say something about how M should be chosen. According to Efron and Tibshirani (1993, Monographs on Statistics Applied Probability 57), 200 bootstrap replications are usually enough for obtaining reasonable estimates of the standard error. Could optimizing the number of bootstrap replications be a possible solution to some of the computational problems pointed out by the authors?

Although it is certainly true that M=10,000 bootstrap replications is excessive, 200 may in some cases be on the low side. We found in our global study of return values for marine wind and significant wave height (Breivik et al, 2014, supplementary figure 7) that the confidence intervals tend to stabilize around 500 bootstrap replications when we look at GPD return estimates. We have chosen a very high number of bootstrap replications here for no better reason than because we could afford it, and because for some tail parameters it is desirable. We have included a new Fig 6 which shows the convergence of the CIs as a function of the number of bootstrap resamples from 50 to 10,000 for non-parametric in-sample estimates of the 100-year return value for significant wave height. The figure shows that indeed for the data set considered we can settle for 1,000 or perhaps slightly fewer bootstrap replications, but probably not as little as 200. To go with Fig 6 we include the following text (page 5, *lines 4-9):*

“It is also of interest to investigate just how many bootstrap resamples are actually needed to obtain CIs from a non-parametric bootstrap technique. In Fig 5 we chose M = 10,000. As Fig 6 shows, this is clearly excessive for reasonable thresholds K0. In fact, Efron and Tibshirani (1993) state that 200 resamples are normally enough.We find this to be on the low side in our case, as Fig 6 shows. However, 1000 resamples is sufficient in this case, but this should be investigated in each case. Breivik et al. (2014) found (see their Supplementary Fig 7) that for a similar data set, 500 resamples would be sufficient when employing a Generalized Pareto Distribution (GPD) on threshold exceedances.”

Specific comments:

Page 1, Line 3: "confidence intervals . . . can be estimated". I would replace "estimated" with "obtained" everywhere, since the intervals are random variables and not parameters.

Agreed.

Page 1, Line 13: In the light of my Major Comment 1, I would not say that ?This is a straightforward procedure?; it is not the computational or algorithmic aspects of a method that matter most, but its validity.

We agree, and we have rewritten the Introduction to emphasize that the procedure is straightforward, but the method of non-parametric bootstrapping has been found to lead to too narrow CIs (low coverage), see page 1, line 13-14. We have also included the following text in the Discussion (page 7, lines 5-15): “As mentioned in the Introduction, an important question is whether non-parametric bootstraps yield CIs with sufficient coverage, ie, CIs that are wide enough. This has been extensively studied by Kysely (2008) who found that non-parametric bootstraps in particular, but also parametric bootstraps tend to have too low coverage. This problem is not addressed by our study, and it is clear that alternative methods are often called for. In particular, the Test Inversion Bootstrap (Carpenter, 1999) is a promising method where the test inversion refers to the duality between hypothesis 10 testing and confidence intervals. Schendel and Thongwichian (2015, 2017) show how this method, originally developed for estimation of statistics of single parameters in the presence of nuisance parameters, can be extended to handle return levels which depend on three parameters for both the Generalized Extreme Value Distribution and GPD by utilizing a maximum likelihood technique. However, non-parametric bootstraps represent a quick and hypothesis-free approach to obtaining CIs, and as the results presented show we can comfortably assume that the results will remain unchanged if we select a small subset of the original sample, provided we follow the procedure outlined in Section 2.??

Below is a full account of the differences between this version and the previous version of the manuscript.

Yours sincerely,
Øyvind Breivik and Ole Johan Aarnes,
Bergen, 2017-02-21

Efficient Bootstrap Estimates for Tail Statistics

Øyvind Breivik[1,2] and Ole Johan Aarnes[1]

[1]Norwegian Meteorological Institute
[2]Geophysical Institute, University of Bergen

Correspondence to: Øyvind Breivik, Norwegian Meteorological Institute, Allegaten 70, NO-5007 Bergen, Norway. E-mail: oyvind.breivik@met.no. ORCID Author ID: 0000-0002-2900-8458

Abstract. Bootstrap resamples can be used to investigate the tail of empirical distributions as well as return value estimates ~~based on~~ from the extremal behaviour of the ~~distribution~~sample. Specifically, the confidence intervals on return value estimates or bounds on in-sample tail statistics can be ~~estimated~~ obtained using bootstrap techniques. However, non-parametric bootstrapping from the entire ~~data set~~ sample is expensive. It is shown here that it suffices to bootstrap from a small subset consisting of the highest entries in the sequence to make estimates that are essentially identical to bootstraps from the entire ~~sequence~~sample. Similarly, bootstrap estimates of confidence intervals of threshold return estimates are found to be well approximated by using a subset consisting of the highest entries. This has practical consequences in fields such as meteorology, oceanography and hydrology where return ~~estimates are routinely made~~ values are calculated from very large gridded model integrations spanning decades at high temporal resolution or from large ensembles of independent and identically distributed model fields. In such cases the computational savings are substantial.

1 Introduction

Bootstrap resamples of time series are commonly used to ~~estimate confidence intervals~~ obtain non-parametric confidence intervals (CIs) on return values (Naess and Clausen, 2001; Naess and Hungnes, 2002) and to investigate the behaviour of the tail of the empirical distribution (Coles, 2001; Beirlant et al., 2006; Qi, 2008). ~~This is a straightforward procedure, but one which~~ Although non-parametric CIs tend to be too narrow, see Kyselý (2008) , the procedure itself is algorithmically and numerically straightforward to implement and is thus a convenient technique for rapidly assessing the width of CIs without having to assume a certain parametric distribution. However, this approach quickly becomes cumbersome for large data sets as it demands random draws from the entire ~~data set~~ sample which subsequently must be sorted to get to the upper percentiles. When handling long model integrations in meteorology, hydrology and oceanography with spatially gridded fields of typically 10^6 grid points this brute-force approach becomes impractical. Such quantities are regularly encountered when estimating return levels from atmospheric reanalyses (Kalnay et al., 1996; Saha et al., 2010; Compo et al., 2011; Dee et al., 2011; Poli et al., 2016), wave hindcasts (Swail and Cox, 2000; Caires and Sterl, 2005; Gaslikova and Weisse, 2006; Breivik et al., 2009; Reistad et al., 2011; Aarnes et al., 2012) and long climate integrations that cover decades or even centuries (Hersbach et al., 2015). When even larger data sets are used, such as the ensembles of seasonal integrations (Stockdale et al., 2011; Molteni et al., 2011), as was done by Van den Brink et al. (2005) on a ~~data set~~ sample amounting to nearly 1,000 years, ~~the data~~

~~processing becomes nearly intractable and~~ finding ways to reduce the size of the ~~data sets~~ samples becomes essential. That is the subject of this paper.

We will present a simple argument for why it is sufficient to retain only a small subset K_0 consisting of the highest entries in a ~~data set~~ sample when estimating tail statistics such as return levels and their associated ~~confidence intervals~~ CIs by means of non-parametric bootstrapping. These highest entries will normally only represent a small fraction of the total ~~data set~~sample. This reduces the need for sorting and storage by several orders of magnitude. The method also reduces the task of sorting the original ~~data set~~ sample as only the K_0 highest entries are kept.

This paper is organised as follows. Sec 2 presents the binomial argument for why we can bootstrap from a small subset consisting of the highest entries in the original ~~data set~~sample. Sec 3 presents three examples of bootstrapped ~~confidence intervals~~ CIs of various tail statistics for a data set of significant wave height from the central North Sea. Here we also show how the method laid out in Sec 2 can be used in practice to determine how many entries must be kept in order to perform an unbiased bootstrap. Sec 4 summarises the results and presents the conclusions.

2 Bootstrapping from the K_0 highest entries in a ~~data set~~sample

Consider the ~~sequence~~ sample $\mathcal{D}_0$ of independent and identically distributed (iid) random numbers $X_1, X_2, \dots, X_N$. Let $X_{N,1} \le X_{N,2} \le \dots \le X_{N,N}$ denote the order statistics on $\mathcal{D}_0$. When investigating a statistic θ which is a function of the k highest entries in $\mathcal{D}_0$, ie $\theta = f(X_{N,N-k+1}, X_{N,N-k+2}, \dots, X_{N,N})$, it is common to form M bootstrap resamples $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_M$, each of length N (Diaconis and Efron, 1983; Efron and Gong, 1983). This method can be used to compute the ~~confidence intervals~~ CIs around extreme value estimates (Breivik et al., 2013, 2014). The procedure is computationally intensive and memory-consuming, as it involves bootstrapping and storing $M \times N$ numbers and performing M sorts, each a process of ~~$\mathcal{O}(N \log N)$ operations~~ $\mathcal{O}(N \log_2 N)$ operations (Press et al. 2007 , pp 423–427). Since we are only interested in combinations of the k highest entries in the resamples $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_M$, we will explore the possibility of instead resampling from only the highest ~~$X_{N,N-K_0+1}, X_{N,N-K_0+2}, \dots, X_{N,N-k+1}, \dots, X_{N,N}$~~ $X_{N,N-K_0+1}, X_{N,N-K_0+2}, \dots, X_{N,N-k+1}, \dots, X_{N,N}$ entries in $\mathcal{D}_0$ ($K_0 > k$). This will be referred to as the *resample threshold* and is sometimes more conveniently written as the percentage of data left out, $P_0 = 100(1 - K_0/N)$.

The probability of drawing one of the highest K_0 entries in $\mathcal{D}_0$ is a binomial problem with probability $p = K_0/N$. The probability of making exactly k draws (with replacement) from the highest K_0 in N draws is thus given by the binomial probability mass function [Zwillinger 1996, p 581]

$$f_{\mathrm{binom}}(k; N, p) = \mathrm{P}(X = k) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{1}$$

where X is a random variable representing the number of draws. The probability of drawing fewer than the required k entries from the highest K_0 is given by the binomial cumulative distribution function

$$F_{\mathrm{binom}}(k-1; N, p) = \mathrm{P}(X < k) = \sum_{i=0}^{k-1} \binom{N}{i} p^i (1-p)^{N-i}. \tag{2}$$

A full bootstrap resample $\mathcal{D}_i$ of length N from $\mathcal{D}_0$ will contain K_i entries from the highest K_0, and $K_i \sim \mathrm{Binom}(N,p)$ where $E[K_i] = K_0$ since the expected (mean) value of the binomial distribution (1) is

$$\mu_{\mathrm{binom}} = Np = K_0. \tag{3}$$

The variance is

$$\sigma^2_{\mathrm{binom}} = Np(1-p) = K_0 - K_0^2/N \approx K_0 \text{ when } K_0 \ll N. \tag{4}$$

Denote a short bootstrap resample from the K_0 highest entries in $\mathcal{D}_0$ as $\tilde{\mathcal{D}}_i$. Two conditions must be met for $\tilde{\mathcal{D}}_i$ to be an unbiased substitute for $\mathcal{D}_i$:

1. The number K_0 must be set large enough that the probability that we miss entries smaller than $X_{N,N-K_0+1}$ in $\mathcal{D}_0$ is below a chosen threshold p_{c}.
2. The length $\tilde{K}_i$ of $\tilde{\mathcal{D}}_i$ must have the same mean and variance as K_i (Eqs 3–4).

To ~~fulfil~~ fulfill Condition (1) it is sufficient to decide on an acceptable level for p_{c}. This probability can be found by consulting Eq (2). It is important to note that choosing K_0 too small will bias the statistic $\tilde{\theta} = f(\tilde{\mathcal{D}}_i)$ since it will be estimated from bootstrap samples that miss entries smaller than $X_{N,N-K_0+1}$. We will for this reason refer to p_{c} as the *probability of contamination* as it gives the probability that the bootstrap estimate is biased because we have kept too few entries from the original ~~data set~~ sample $\mathcal{D}_0$. A very conservative bound on p, and thus on $K_0 = Np$, can be found quickly by consulting Hoeffding's formula (Hoeffding, 1963),

$$F(k;N,p) \leq \exp\left(-2\frac{(Np-k)^2}{N}\right), \tag{5}$$

valid when $k \leq Np$. A useful quantity is the ratio $r = K_0/k$ of upper entries retained (K_0) and the minimum number k required to form a bootstrap estimate of the statistic in question for a given probability of contamination p_{c}. This can be ~~estimated~~ found from Eq (2), but when N is large the Poisson distribution is a good approximation and more practical to work with,

$$F_{\mathrm{Poisson}}(k-1;rk) = \mathrm{P}(X<k) = \mathrm{e}^{-rk}\sum_{i=0}^{k-1}\frac{(rk)^i}{i!}. \tag{6}$$

Fig 1 shows the minimum acceptable ratio K_0/k as a function of k for levels of p_{c} ranging from 10^{-5} to 0.05. The probabilities can be computed from Eq (2) [or more conveniently from Eq (6)]. As can be seen, for all values of k, the ratio is comfortably below 15, and for values of k larger than 10 a ratio of 3 is sufficient even for a confidence level of 10^{-5}. See the appendix for a more detailed explanation of the ratio curves used throughout.

Condition (2) can be handled by randomly perturbing the size of the resamples, $\tilde{K}_i$, such that it mimics the number of draws, $K_i \sim \mathrm{Binom}(N,p)$, that would have been made from the upper K_0 entries of $\mathcal{D}_0$ in a full bootstrap $\mathcal{D}_i$. In practice, as we shall see, the statistics are quite insensitive to these perturbations as long as K_0 has been chosen sufficiently large.

3 Bootstrapping confidence intervals

Here we present worked examples of how the two conditions presented above can reduce the problem of estimating ~~confidence intervals~~ CIs on tail statistics for a data set of independent ensemble forecasts at long lead time ($N = 330{,}000$). We use archived ensemble forecasts (Molteni et al., 1996) of significant wave height in the central North Sea (near the Ekofisk oil field at 56.5° N, 003.2° E; a histogram of the data set used is shown in Fig 2) at a forecast lead time of 240 hours. 100-year return values from these ensembles have previously been reported by ~~Breivik et al. (2013) and Breivik et al. (2014)~~ Breivik et al. (2013, 2014) .

3.1 Example 1: Confidence intervals on in-sample return estimates

Consider as an example the problem of how to calculate in-sample return estimates from the ~~data set~~ sample of independent forecasts presented above. These forecasts can be considered iid (as they are not from correlated time series). An in-sample return estimate is calculated directly from the tail of the empirical distribution rather than by applying extreme value analysis. As explained by Breivik et al. (2013) the independent forecasts presented in Fig 2 add up to the equivalent of 229 years under the assumption that each forecast represents a time interval $\Delta t = 6$ hours. A 100-year return estimate is then a linear interpolation between $X_{N,N-1}$ and $X_{N,N-2}$ (the second and third highest entries in $\mathcal{D}_0$),

$$H_{100} = 0.67X_{N,N-1} + 0.33X_{N,N-2}. \tag{7}$$

Now, clearly $k = 3$ since we need the second and third highest entries in our resamples to form a return estimate. Let us now tentatively keep the $K_0 = 1,000$ highest entries and bootstrap from these instead of from the entire sequence to compute the ~~confidence intervals~~ CIs on the linear combination of the second and third highest entries given by Eq (7). The size $\tilde{K}_i$ of the resamples, $\tilde{\mathcal{D}}_i$, is drawn from the binomial distribution (Eqs 3–4) with $\mu = K_0$ and $\sigma^2 \approx K_0$. What is the probability p_c that one of the three highest entries in a bootstrapped sequence should *not* have come from the 1,000 highest entries that we have retained (i.e. should depend on entries contained in the bulk of the ~~data set~~ sample that we discarded)? It is clear that the probability of drawing one of the highest 1,000 entries is $p = 1{,}000/330{,}000$, and from Eq (2) we find that the probability of picking too few (< 3) entries from the K_0 highest is

$$F(2; 330,000, p) = \mathrm{P}(X \leq 2) = \sum_{i=0}^{2} \binom{330{,}000}{i} p^i (1-p)^{330{,}000-i}, \tag{8}$$

which is indistinguishable from zero to double precision. Reducing the number K_0 to 10 ($r \approx 3$) raises the probability of contaminating the resamples by entries from the lower $N - K_0$ to 0.002. This can also be confirmed by consulting Fig 1 for the combination $k = 3, r = 3$. For $M = 1{,}000$ resamples we may thus expect on average 2 resamples to be contaminated by values from the lower $N - K_0$ values in the original sequence. A very safe compromise in this case is $K_0 = 100$ ($r \approx 33$). Consulting Fig 1 shows that for $k = 3, r = 33$ we are well below a probability of contamination of 10^{-5}. The quantile-quantile (QQ) plot in Fig 3 shows that resampled return estimates of significant wave height from the full ~~data set~~ sample $\mathcal{D}_0$ (see Fig 2) have practically the same distribution as resamples from the upper $K_0 = 100$ entries.

Condition (2) given above states that the ~~length~~ size of the reduced resamples $\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_M$ should be randomly perturbed around the mean value K_0. In practice this condition turns out to be rather insignificant as long as K_0 is chosen sufficiently large. This is demonstrated in the QQ plot in Fig 4 where we see that perturbed-length estimates (abscissa) closely match the distribution of fixed-length estimates (ordinate). However, choosing K_0 too small will bias the statistic in question. This is illustrated in Fig 5 where we see that bootstrap estimates from too-short ~~data sets~~ subsets of the original sample (K_0 chosen too small) are biased high. As K_0 approaches 30 ($r = 10$), the mean and standard deviation of the return estimates approach their asymptotic values. These findings are in accordance with what we find by consulting Fig 1 where we see that $k = 3, r = 10$ has a probability of contamination p_{c} less than 10^{-5}. It is also of interest to investigate just how many bootstrap resamples are actually needed to obtain CIs from a non-parametric bootstrap technique. In Fig 5 we chose $M = 10,000$. As Fig 6 shows, this is clearly excessive for reasonable thresholds K_0. In fact, Efron and Tibshirani (1993) state that 200 resamples are normally enough. We find this to be on the low side in our case, as Fig 6 shows. However, 1000 resamples is sufficient in this case, but this should be investigated in each case. Breivik et al. (2014) found (see their Supplementary Fig 7) that for a similar data set, 500 resamples would be sufficient when employing a Generalized Pareto Distribution (GPD) on threshold exceedances.

3.2 Example 2: Confidence intervals on upper percentiles

A similar problem to the estimation of ~~confidence intervals~~ CIs for in-sample return values is how to ~~estimate the confidence interval~~ obtain the CI for the highest percentiles, e.g. the 99th percentile (P_{99}). The upper percentile is frequently used when investigating trends in for example the wind and wave height climate [see e.g. Wang and Swail (2001, 2002)]. In order to construct a bootstrap estimate of P_{99} brute force it is necessary to resample the entire ~~data set~~ sample $\mathcal{D}_0$ and sort the bootstrap to get to the $N/100$-th highest entry. However, Fig 1 tells us that when $k = N/100$ is large (as it will be when N is large), we can with extremely high certainty say that keeping the $K_0 = 2k$ highest entries is enough to perform a bootstrap resample exercise for the ~~confidence interval~~ CI on P_{99}. In fact, $K_0 = 1.2k$ is sufficient for all significance levels plotted in Fig 1. This means that in order to ~~estimate a confidence interval~~ obtain a CI for P_{99} we need only find the entry $X_{N,N-k}$ that corresponds to P_{99} from the original ~~data set~~ sample $\mathcal{D}_0$ and retain entries higher than $X_{N,N-1.2k}$. Fig 7 shows how the ratio r decreases as the sample size N increases. It is clear that for all probabilities of contamination investigated, a ratio of $K_0/k = 2$ is sufficient when N is larger than 2,000. Obviously, samples smaller than $\mathcal{O}(10^3)$ do not pose computationally demanding problems anyway and are of no interest to us in this context. Fig 8 illustrates for a fixed probability of contamination $p_{\mathrm{c}} = 0.01$ that even as we go to higher percentiles (the uppermost curve shows $P_{99.9}$), a ratio $K_0/k = 2$ is sufficient as the sample size N exceeds 10^4 (see the appendix for more details on the ratio curves).

3.3 Example 3: Confidence intervals on return estimates from threshold exceedances

Consider now the problem of estimating ~~confidence intervals for threshold exceedances. The Generalized Pareto distribution (GPD)~~ CIs for return estimates from threshold exceedances from a data set of *independent forecasts*. This differs from a peaks-over-threshold approach which is how correlated time series must be handled to estimate return levels (Coles, 2001) .

GPD gives the relevant extreme value distribution for independent exceedances above a threshold u (Coles 2001, pp 75–77),

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}. \tag{9}$$

Here $y = X_i - u$, $y > 0$ are exceedances above ~~an entry X_k,~~ a threshold $u = X_{N,N-k+1}$ (remember that $X_{N,N-k+1}$ is the k-th highest entry in the sample $\mathcal{D}_0$) and $\tilde{\sigma}$ is a scale parameter which is a function of the threshold u, and ξ is the shape parameter.

A brute-force approach would be to make N draws from $\mathcal{D}_0$ ~~–~~(with replacement) and repeat this procedure M times. Then, GPD return estimates would be computed for each of the resulting bootstrap sequences $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M$. Say we want to try to instead ~~keep~~ retain only the K_0 entries exceeding a threshold U_0, where $U_0 < u$, corresponding to the entry ~~X_{K_0}~~ $X_{N,N-K_0+1}$ in the original ~~data set~~ sample $\mathcal{D}_0$. From these we need to draw at least k entries, from which we will make return estimates. The question is again how many entries (K_0) must be kept to arrive at an acceptably low probability p_c that the statistic should really be based on entries below the threshold U_0.

This problem arises when estimating GPD return values from the independent ensemble forecasts (Fig 2). For such a ~~data set~~ sample all exceedances above a given threshold can be used to form GPD return value estimates (9). ~~Confidence intervals~~ CIs on the return values can likewise be ~~estimated~~ obtained by bootstrapping from all entries exceeding this threshold. For a large ~~data set~~ sample this is orders of magnitude faster than bootstrapping from the entire ~~data set~~sample. Assume again that we have kept all forecasts exceeding $P_{99.1}$, ie the $K_0 = 3{,}000$ highest entries (cf Fig 2). To form a return estimate we assume that we need at least $k = 1{,}000$ entries, corresponding to $P_{99.7}$. The probability of drawing (with replacement) k or fewer entries from the highest K_0 in N draws can again be found from Eq (2) and is indistinguishable from zero to double precision with the given choice of N, K_0 and k. This is easy to verify by consulting Fig 1 where we see that for $k = 1000, r = 3$ we are well above the 10^{-5} level. Fig 9 shows that the confidence interval and the mean return value based on $M = 1{,}000$ bootstrap resamples for various choices of resample threshold $100(1 - K_0/N)$ (i.e. the percentage of data omitted) are practically identical to the ~~confidence intervals based on the full data set~~ CIs based on $\mathcal{D}_0$ (marked as asterisks). Only when $r = K_0/k$ comes close to unity do we experience fluctuations and biases (i.e., the resample threshold nearly coincides with the number of tail entries required to form a return estimate, in this case the threshold $P_{99.7}$).

4 Conclusions

~~Confidence intervals~~ CIs and other statistics ~~on~~ of the extremes and the tail of empirical distributions are commonly found using non-parametric bootstrap techniques. Here we have shown that it is unnecessary to bootstrap from the entire ~~data set~~original sample. The actual number K_0 highest entries that must be kept to make unbiased bootstrap estimates for the tail of an empirical distribution depends on $K_0 = Np$ as well as on the number k highest entries that are required for the statistic in question. The examples in the previous sections calculated p_c given a predetermined number K_0 of tail entries that have been kept. This is a realistic approach as in practice we often retain a larger part of the tail of an empirical distribution than what is strictly needed since the same data set is used to compute other statistics. It is then sufficient to consult Eq (2) to determine whether K_0 is sufficiently large. A quick estimate of the probability of contamination can be made by consulting Fig 1.

The advantages of restricting resamples to a small subset K_0 consisting of the highest entries in $\mathcal{D}_0$ can be summarised as follows. First, only the upper K_0 entries need be kept and sorted in the original data set. This offers substantial savings in cases like those described by Breivik et al. (2013, 2014) where a very large number of forecasts ($>$300,000) are handled, each consisting of more than 60,000 grid points in space. Second, the size of the resamples is also reduced from N to an average size K_0, where K_0 is usually a very small fraction of N, typically less than 1%. Third, this reduction in resample size also means that the cost of sorting the resamples to get to the highest entries is greatly reduced, as the problem is now linear in the number of bootstrap resamples M since each sort is $\mathcal{O}(K_0 \log_2 K_0)$, which is now a constant number independent of the size of the original sample (or a small fraction of it, as in the 99th percentile shown in Example 2).

We have investigated the conditions that must be met to form a non-parametric bootstrap for tail statistics such as return levels (which depend on all three parameters of the Generalized Extreme Value Distribution or the GPD) from a small subset of the highest entries in the original sample. As mentioned in the Introduction, an important question is whether non-parametric bootstraps yield CIs with sufficient coverage, ie, CIs that are wide enough. This has been extensively studied by Kyselý (2008) who found that non-parametric bootstraps in particular, but also parametric bootstraps tend to have too low coverage. This problem is not addressed by our study, and it is clear that alternative methods are often called for. In particular, the Test Inversion Bootstrap (Carpenter, 1999) is a promising method where the test inversion refers to the duality between hypothesis testing and confidence intervals. Schendel and Thongwichian (2015, 2017) show how this method, originally developed for estimation of statistics of single parameters in the presence of nuisance parameters, can be extended to handle return levels which depend on three parameters for both the Generalized Extreme Value Distribution and GPD by utilizing a maximum likelihood technique. However, non-parametric bootstraps represent a quick and hypothesis-free approach to obtaining CIs, and as the results presented show we can comfortably assume that the results will remain unchanged if we select a small subset of the original sample, provided we follow the procedure outlined in Section 2.

Appendix: Consulting the ratio curves

The ratio curves presented in Figs 1, 7 and 8 are convenient for quickly establishing how many entries (K_0) must be kept in order to form an unbiased resample that depends on the highest k entries. The relationship between Fig 1 and Fig 7 can be illustrated as follows. If we assume N large we can use Fig 1. In practice we can choose $N = 2 \times 10^3$ without violating the assumption that N is large. Now assume that the statistic in question is the 99th percentile, i.e. $k = N/100 = 20$. Let us choose a probability of contamination $p_\mathrm{c} = 0.01$ (this corresponds to the red curve marked with diamonds in Fig 1). We find the ratio to be 1.6, i.e. we will need to keep 60% more entries than the entry corresponding to P_{99}. The corresponding curve in Fig 7 is also marked in red. Here, the location on the x-axis to read off is $N = 2 \times 10^3$ which lies on the y-axis, and the ratio is again found to be 1.6. A more realistic example in terms of sample size would be $N = 10^5$ (and $k = N/100 = 10^3$). Now we find from either Fig 1 or Fig 7 that with a probability of contamination $p_\mathrm{c} = 0.01$ that the ratio is 1.13, i.e. we need only keep 13% more entries than the one representing the 99 percentile. Figs 1, 7 and 8 clearly illustrate that in almost all cases it is sufficient to retain at most twice as many entries K_0 from the tail of the sample distribution $\mathcal{D}_0$ than what is required (k) for the statistic in question.

Acknowledgements. This study was carried out with support from the Research Council of Norway through the ExWaCli project (grant no 226239) and the follow-up, ExWaMar. The data sets presented in this study are archived in the MARS database of the European Centre for Medium-Range Weather Forecasts (ECMWF), see http://old.ecmwf.int/services/archive/.

References

Aarnes, O. J., Breivik, Ø., and Reistad, M.: Wave Extremes in the Northeast Atlantic, J Climate, 25, 1529–1543, doi:10/bvbr7k, 2012.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J.: Statistics of extremes: theory and applications, John Wiley & Sons, 2006.

Breivik, Ø., Gusdal, Y., Furevik, B. R., Aarnes, O. J., and Reistad, M.: Nearshore wave forecasting and hindcasting by dynamical and statistical downscaling, J Marine Syst, 78, S235–S243, arXiv:1206.3055, doi:10/cbgwqd, 2009.

Breivik, Ø., Aarnes, O. J., Bidlot, J.-R., Carrasco, A., and Saetra, Ø.: Wave Extremes in the Northeast Atlantic from Ensemble Forecasts, J Climate, 26, 7525–7540, arXiv:1304.1354, doi:10/mpf, 2013.

Breivik, Ø., Aarnes, O., Abadalla, S., Bidlot, J.-R., and Janssen, P.: Wind and Wave Extremes over the World Oceans From Very Large Ensembles, Geophys Res Lett, 41, 5122–5131, arXiv:1407.5581, doi:10.1002/2014GL060997, 2014.

Caires, S. and Sterl, A.: A new nonparametric method to correct model data: application to significant wave height from the ERA-40 re-analysis, J Atmos Ocean Tech, 22, 443–459, doi:10.1175/JTECH1707.1, 2005.

Carpenter, J.: Test inversion bootstrap confidence intervals, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61, 159–172, doi:10.1111/1467-9868.00169, 1999.

Coles, S.: An introduction to statistical modeling of extreme values, Springer Verlag, 2001.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The Twentieth Century Reanalysis Project, Q J R Meteorol Soc, 137, 1–28, doi:10.1002/qj.776, 2011.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., P, B., Beljaars, A., van de Berg, L., Bidlot, J., Bormann, N., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q J R Meteorol Soc, 137, 553–597, doi:10.1002/qj.828, 2011.

Diaconis, P. and Efron, B.: Computer intensive methods in statistics, Scientific American, 248, 116–130, 1983.

Efron, B. and Gong, G.: A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, The American Statistician, 37, 36–48, 1983.

Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, CRC press, 1993.

Gaslikova, L. and Weisse, R.: Estimating near-shore wave statistics from regional hindcasts using downscaling techniques, Ocean Dyn, 56, 26–35, 2006.

Hersbach, H., Peubey, C., Simmons, A., Berrisford, P., Poli, P., and Dee, D.: ERA-20CM: a twentieth-century atmospheric model ensemble, Q J R Meteorol Soc, 141, 2350–2375, doi:10.1002/qj.2528, 2015.

Hoeffding, W.: Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association, 58, 13–30, doi:10.1080/01621459.1963.10500830, 1963.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al.: The NCEP/NCAR 40-Year Reanalysis Project, Bull Am Meteor Soc, 77, 437–472, doi:10/fg6rf9, 1996.

Kyselý, J.: A Cautionary Note on the Use of Nonparametric Bootstrap for Estimating Uncertainties in Extreme-Value Models, Journal of Applied Meteorology and Climatology, 47, 3236–3251, doi:10.1175/2008JAMC1763.1, 2008.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Q J R Meteorol Soc, 122, 73–119, doi:10.1002/qj.49712252905, 1996.

Molteni, F., Stockdale, T., Balmaseda, M. A., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T. N., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), ECMWF Technical Memorandum 656, European Centre for Medium-Range Weather Forecasts, 2011.

Naess, A. and Clausen, P.: Combination of the Peaks-over-treshold and bootstrapping methods for extreme value prediction, Structural Safety, 23, 315–330, 2001.

Naess, A. and Hungnes, B.: Estimating Confidence Intervals of Long Return Period Design Values by Bootstrapping, Journal of Offshore Mechanics and Arctic Engineering, 124, 5, doi:10.1115/1.1446078, 2002.

Poli, P., Hersbach, H., Dee, D., Berrisford, P., Simmons, A., Vitart, F., Laloyaux, P., Tan, D., Peubey, C., Thepaut, J.-N., Trémolet, Y., Hólm, E., Bonavita, M., Isaksen, L., and Fisher, M.: ERA-20C: An Atmospheric Reanalysis of the Twentieth Century, J Climate, 29, 4083–4097, doi:10.1175/JCLI-D-15-0556.1, 2016.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P.: Numerical Recipes in C, 3rd edition, Cambridge University Press, Cambridge, 2007.

Qi, Y.: Bootstrap and empirical likelihood methods in extremes, Extremes, 11, 81–97, doi:10.1007/s10687-007-0049-8, 2008.

Reistad, M., Breivik, Ø., Haakenstad, H., Aarnes, O. J., Furevik, B. R., and Bidlot, J.-R.: A high-resolution hindcast of wind and waves for the North Sea, the Norwegian Sea, and the Barents Sea, J Geophys Res, 116, 18 pp, C05 019, arXiv:1111.0770, doi:10/fmnr2m, 2011.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., et al.: The NCEP climate forecast system reanalysis, Bull Am Meteor Soc, 91, 1015–1057, doi:10.1175/2010Bams3001.1, 2010.

Schendel, T. and Thongwichian, R.: Flood frequency analysis: Confidence interval estimation by test inversion bootstrapping, Advances in Water Resources, 83, 1–9, doi:10.1016/j.advwatres.2015.05.004, 2015.

Schendel, T. and Thongwichian, R.: Confidence intervals for return levels for the peaks-over-threshold approach, Advances in Water Resources, 99, 53–59, doi:10.1016/j.advwatres.2016.11.011, 2017.

Stockdale, T., Anderson, D., Balmaseda, M., Doblas-Reyes, F., Ferranti, L., Mogensen, K., Palmer, T., Molteni, F., and Vitart, F.: ECMWF seasonal forecast system 3 and its prediction of sea surface temperature, Climate Dynamics, pp. 455–471, doi:10.1007/s00382-010-0947-3, 2011.

Swail, V. R. and Cox, A. T.: On the use of NCEP-NCAR reanalysis surface marine wind fields for a long-term North Atlantic wave hindcast, J Atmos Ocean Tech, 17, 532–545, doi:10/dnkmbb, 2000.

Van den Brink, H., Können, G., Opsteegh, J., Van Oldenborgh, G., and Burgers, G.: Estimating return periods of extreme events from ECMWF seasonal forecast ensembles, Int J Climatol, 25, 1345–1354, doi:10.1002/joc.1155, 2005.

Wang, X. and Swail, V.: Changes of extreme wave heights in Northern Hemisphere oceans and related atmospheric circulation regimes, J Climate, 14, 2204–2221, doi:10/dz8fqn, 2001.

Wang, X. and Swail, V.: Trends of Atlantic wave extremes as simulated in a 40-yr wave hindcast using kinematically reanalyzed wind fields, J Climate, 15, 1020–1035, doi:10/fksbwn, 2002.

Zwillinger, D., ed.: CRC Standard Mathematical Table and Formulae, CRC Press, Boca Raton, FL, USA, 1996.

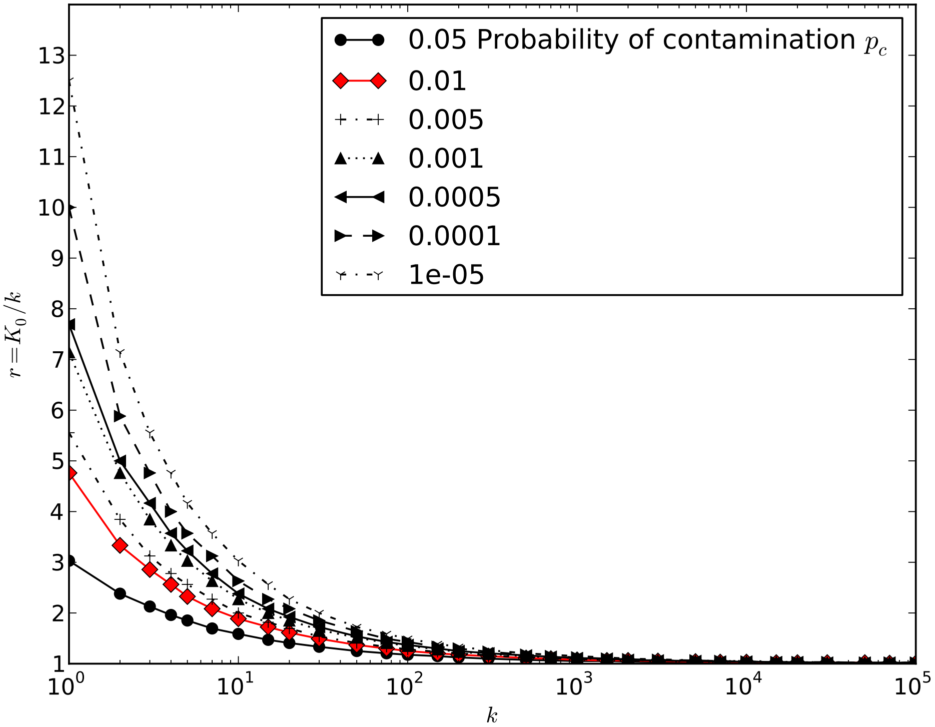

Figure 1. The ratio K_0/k as a function of k, the minimum number of bootstrapped entries needed for the statistic in question, for levels of probability of contamination ranging from 10^{-5} (uppermost curve) to 0.05 (lowermost curve). The curve representing 1% probability of contamination is marked in red (with diamonds) as it is a reasonable confidence level.

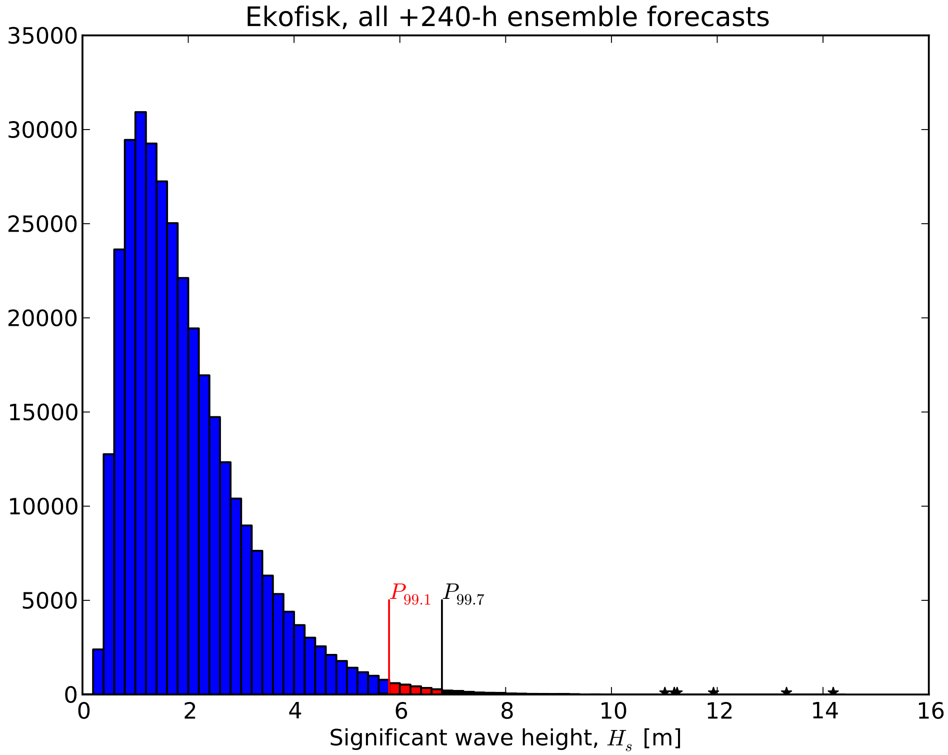

Figure 2. Histogram of the significant wave height from archived ensemble forecasts in the central North Sea (Ekofisk, 56.5N, 003.2E) at +240 h lead time. Entries above $P_{99.1}$, corresponding to threshold U_0, are coloured red whilst entries exceeding $P_{99.7}$, corresponding to the upper threshold, u, are in black. The highest entries are individually marked with asterisks.

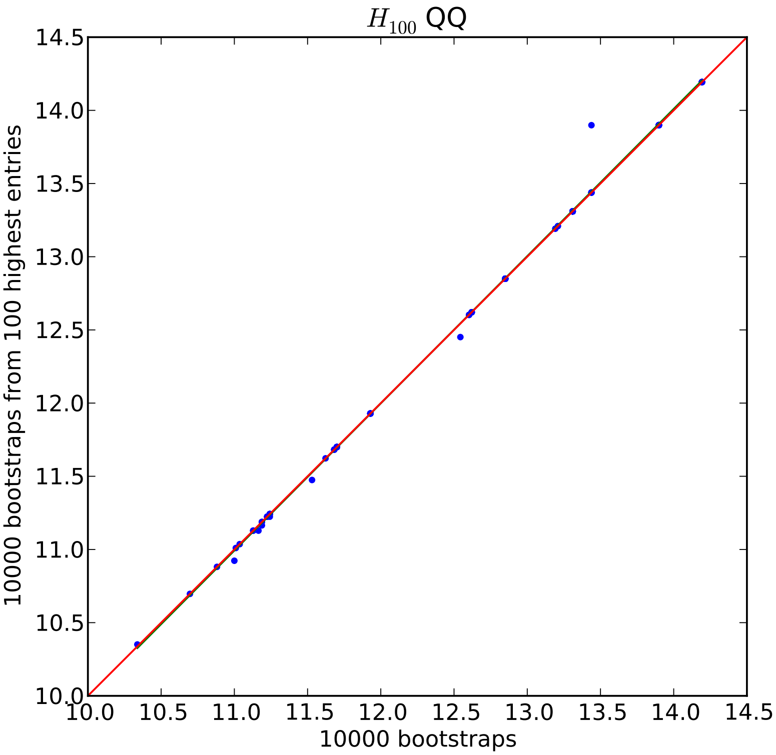

Figure 3. A quantile-quantile comparison of 10,000 bootstrapped direct 100-year return estimates of significant wave height taken from a forecast ensemble (Breivik et al., 2013) versus a bootstrap from the upper 100 entries in the ~~data set~~sample. The 45° line is shown in red.

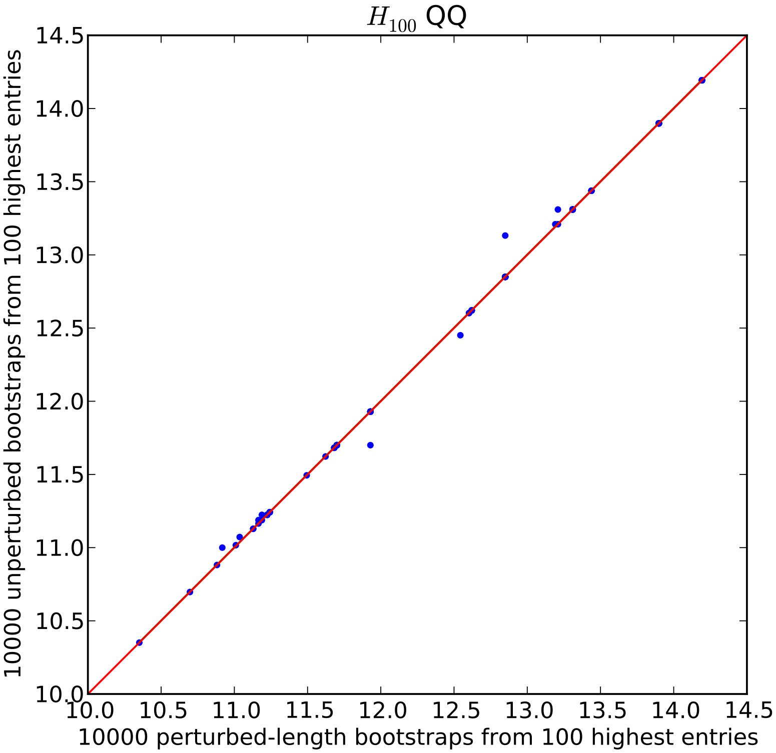

Figure 4. A quantile-quantile (QQ) comparison of $M = 10{,}000$ bootstraps $\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_M$ of variable length $\tilde{K}_1, \tilde{K}_2, \ldots, \tilde{K}_M$ against bootstraps of fixed length K_0, all from the upper 100 entries in the original sequence $\mathcal{D}_0$. The difference is very small.

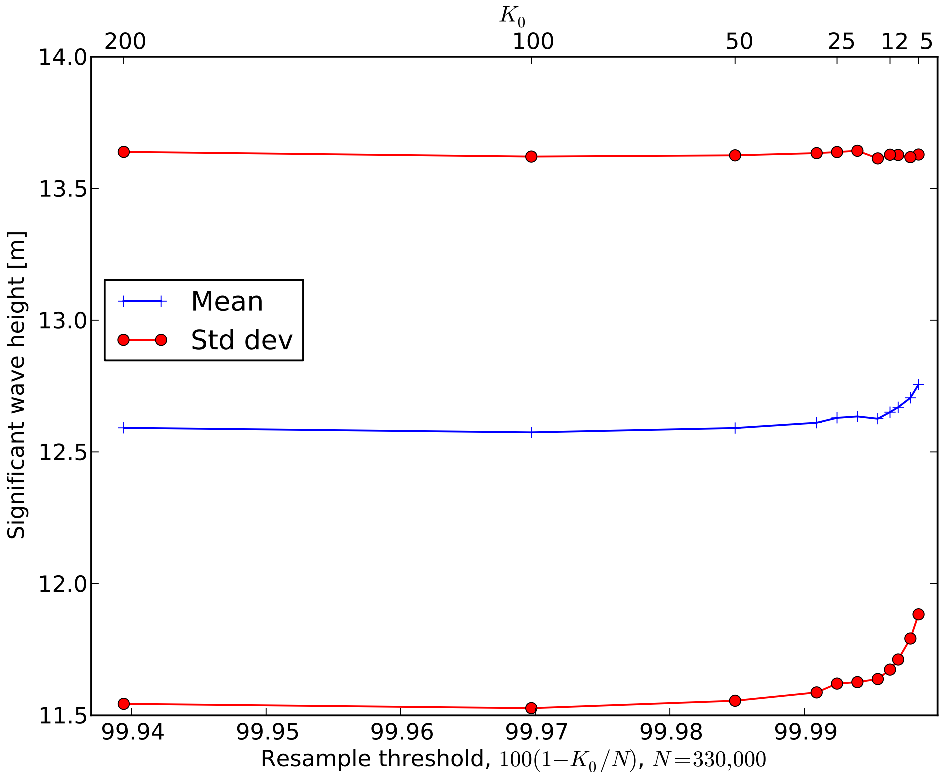

Figure 5. Mean and standard deviation on 100-yr in-sample return estimates based on $M = 10{,}000$ bootstrap resamples for various choices of resample threshold K_0 for the ~~data set~~ sample in Fig 2. A minimum of $k = 3$ entries are required to form the return estimate [see Eq (7)]. For choices of K_0 smaller than 30 (corresponding to a ratio $r = K_0/k = 10$) the bootstrap resamples are biased high.

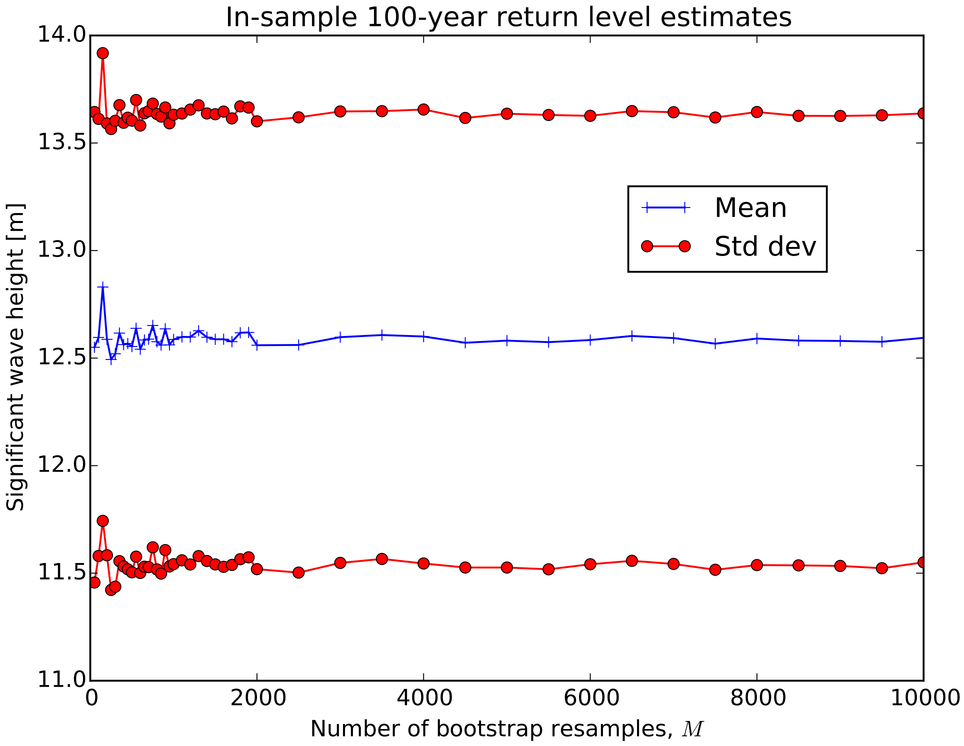

Figure 6. Mean and standard deviation on 100-yr in-sample return estimates with a threshold $K_0 = 1,000$ as a function of number of bootstrap resamples, M. For $M > 1000$ the CIs are quite stable.

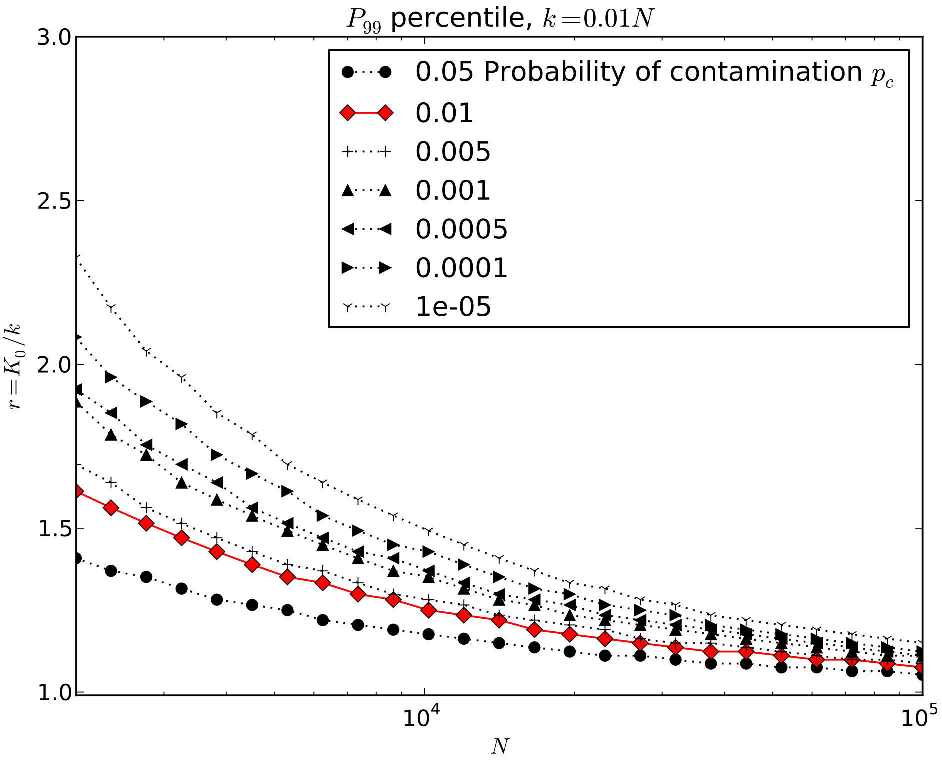

Figure 7. Bootstrapping the 99th percentile, P_{99}. The ratio $r = K_0/k$ is shown as a function of sample size N. Here, the minimum number of entries required is simply the upper 1% (P_{99}), so $k = N/100$. Various levels of probability of contamination p_c are shown, and for sample sizes larger than approximately $2,000$, a ratio $r = 2$ is sufficient. The curve representing 1% probability of contamination is marked in red (with diamonds) as it represents a reasonable confidence level.

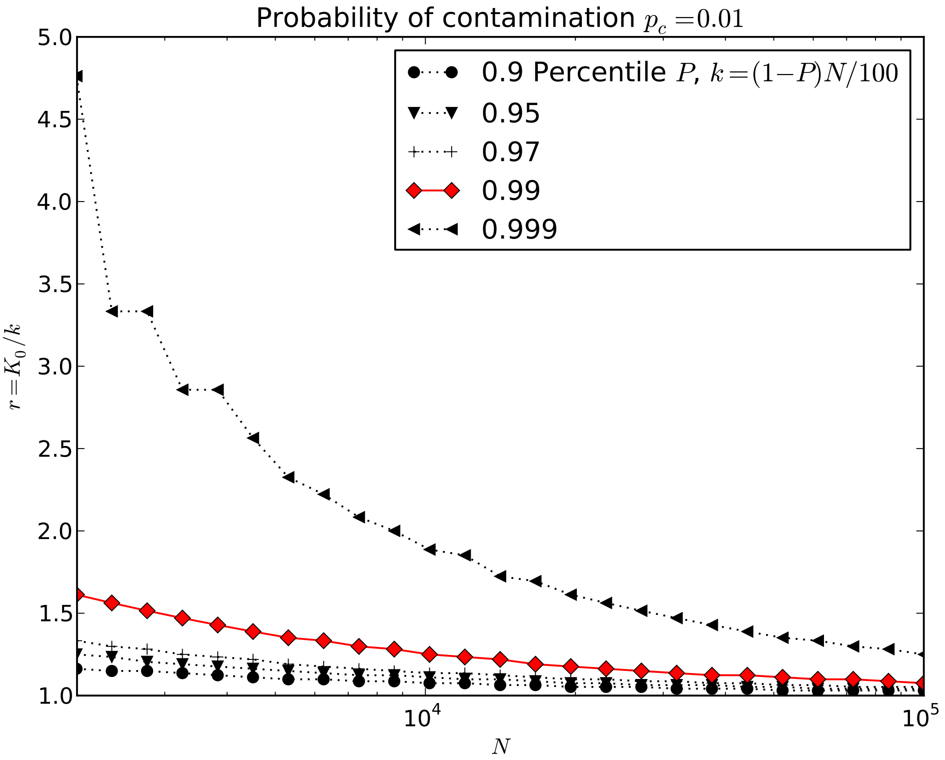

Figure 8. Bootstrapping the upper percentiles $P = P_{90}, P_{95}, P_{97}, P_{99}$ and $P_{99.9}$. The ratio $r = K_0/k$ is shown as a function of sample size N. Here, the minimum number of entries required is $k = (1-P)N/100$. The probability of contamination is kept fixed at $p_c = 0.01$. At sample sizes larger than approximately 10^4, a ratio $r = 2$ is sufficient for all percentiles investigated. The curve representing the 99th percentile is marked in red (with diamonds) and corresponds to the red curve in Fig 7.

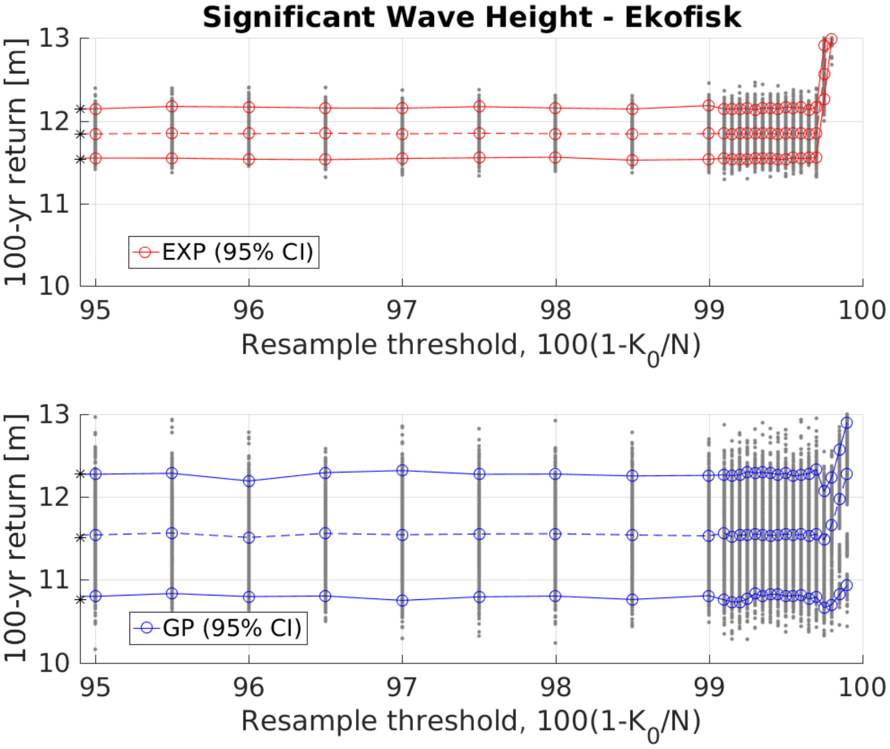

Figure 9. The upper and lower 95% ~~confidence intervals~~ CIs and the mean 100-yr return estimates (dashed) based on $M = 1000$ bootstrap resamples for various choices of resample threshold K_0 for the ~~data set~~ sample in Fig 2. Upper panel: a GPD with shape parameter $\xi = 0$ (exponential distribution). Lower panel: a GPD with freely varying shape parameter. Individual bootstrap estimates are marked in grey. Estimates based on the full ~~data set~~ sample $\mathcal{D}_0$ are marked as asterisks on the ordinate.