Nat. Hazards Earth Syst. Sci. Discuss.,
doi:10.5194/nhess-2016-240-AC2, 2017

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Efficient Bootstrap Estimates for Tail Statistics" *by* Øyvind Breivik and Ole Johan Aarnes

Øyvind Breivik and Ole Johan Aarnes

oyvind.breivik@met.no

Received and published: 17 January 2017

Response to Reviewer 2, Sofia Caires:

Thank you for your thorough review. Our responses are in 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 efficiency 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 will elaborate on this in our revision of Example 3 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 figures 3, 4 and 5 are based on M=10,000 bootstrap replications, while those shown in Figure 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 will include a figure 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 (see below). 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 the figure below we will include the following text:

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 [new] 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 GPD technique.

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 plan to make the appropriate changes to the manuscript by emphasizing that the procedure is straightforward, but the method of non-parametric bootstrapping has been found to lead to too narrow CIs (low coverage). We suggest to incorporate the following text in the discussion:

We have investigated the conditions that must be met to form a non-parametric bootstrap for tail statistics such as return values (which depend on all three parameters of the GEV or GPD). An important question is whether non-parametric bootstraps yields 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 advocated by Schendel and Thongwichian (2015,

2017) is a promising method. However, non-parametric tail statistics are often a necessary first approach to obtaining CIs, and the results presented show that we can comfortably assume that the results will remain unchanged if we take a small subset of the original data set, provided we follow the procedure outlined in Section 2.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., doi:10.5194/nhess-2016-240, 2016.

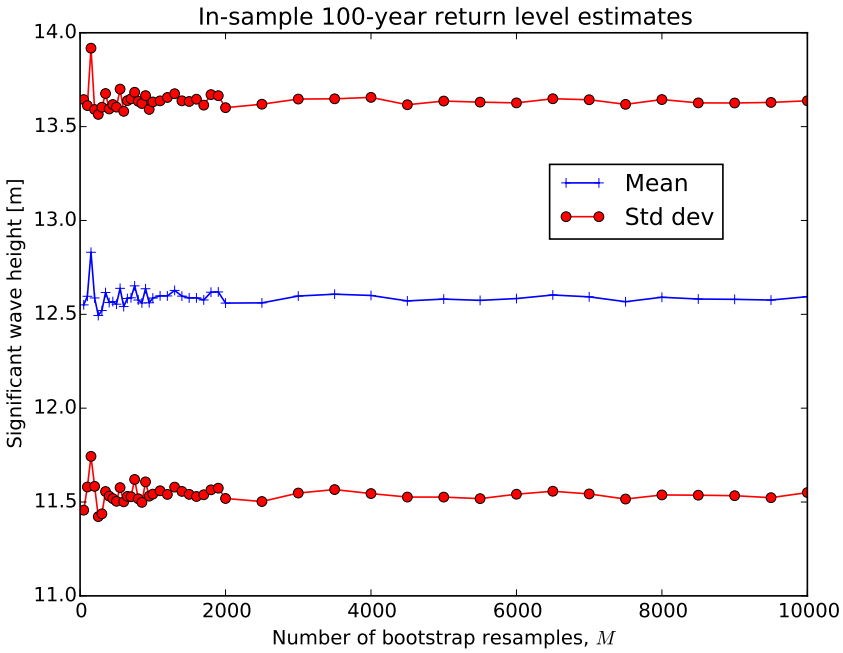

Fig. 1. New Fig 6: Mean and standard dev on 100-yr in-sample return estimates with a threshold K0 = 1,000 as a function of number of bootstrap resamples, M. For M > 1000 the CIs are quite stable.