Estimating return levels from serially dependent extremes

Once again, we thank the Associate Editor and the referees for reviewing our work. We believe the current version of our paper is much improved over earlier versions, and we have the referees and AE to thank for this. In this revision, we

1. use bias–corrected, accelerated bootstrap confidence intervals (Efron, 1987) for return level estimates in the Newlyn sea–surge and Bradfield wind speed data examples (Table 3), instead of the percentile intervals used previously (see Section 4.3 in the curent submission for a more complete description and explanation);
2. have thoroughly checked our standard error calculations for Tables 1 and 2 and are now confident that the values reported here are correct;
3. Add a new section – Section 5 – giving some concluding remarks and spelling out what we believe is 'best practice' when estimating return levels from serially dependent extremes;
4. have thoroughly proof–read our work; as well as correct the minor typographical issues pointed out by the two referees, we have corrected a few other mistakes we have found since the last submission.

We now respond to the comments of the Associate Editor, and both referees, in turn.

Associate Editor's comments

Since the last submission, we have carefully proof-read our manuscript and have corrected the typographical errors pointed out by the two referees, as well as various other mistakes we have found. The Associate Editor said it was not essential for us to address the first two points raised by referee 2 (estimation of θ and model mis–specification). We have, however, investigated the issue of bootstrap confidence intervals (see point 1 above, our comments to referee 2 and Section 4.3 in the most recent version of the paper). We agree with the Associate Editor that the last version of the paper seemed to end somewhat abruptly; we were conscious about the length of the paper and previous requests to reduce this. We have now included a final section (Section 5) which summarises our key findings and spells out what we believe is 'best practice'. We hope the Associate Editor finds this satisfactory.

[Our response to referee 1's report is overleaf]

Referee 1's report

- The reference to Northrop and Jonathan (2011) has now been reinstated (fifth line on page 3 of the most recent submission).
- In Section 2.3.4 (last line of page 6), we now say "see the last row of Table 1", as suggested by the referee.
- The referee points out that the upper bound of our 95% profile (log) likelihood confidence interval for $\hat{z}_{1000}$ changed from 3.365 to 6.452 (page 7, line 2): we should have explained this in our response accompanying the last submission. On re–calculating everything before the last submission, we realised that the upper bound of 3.365 metres is obtained when a cluster separation interval of $\kappa = 10$ observations is used; as our discussion in Section 2.3.4 on page 6 explains, Coles and Tawn (1991) suggest an interval of $\kappa = 20$ for sea surges at this location, and this gives the corrected upper bound of 6.452 metres. The effect of using other values of κ is discussed in the next paragraph (on page 7), and here we explain that the upper bound of 3.365 for $\hat{z}_{1000}$ is obtained when we use $\kappa = 10$. So there was a mistake in our first submission which we corrected in the second submission, without explaining what we had done... sorry!
- As suggested by the referee, we have removed the exclamation mark after 30 (page 7, line 7), just in case readers mistake this as 30 factorial.
- On page 26, we replace "arrivals" with "arrival", as spotted by the referee.
- The referee quite rightly queries our standard errors in Tables 1 and 2. We apologise for the confusion here; we really should have given a deeper explanation in our last response. We do this here.

  (i) In the first submission, we did not attempt to estimate $\mathsf{se}(\hat{\theta}^{[2]})$ and $\mathsf{se}(\hat{\theta}^{[3]})$. In the second submission (and the third!) we do this via a block bootstrap procedure, incorporating this into our estimated standard errors for return levels (via the delta method). The table below reports standard errors for the return levels in the first submission (top row) and the second submission (after quantifying the uncertainty in $\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$):

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
$\hat{\theta}^{[2]}/\hat{\theta}^{[3]}$ constant	(0.091)	(0.127)	(0.201)
$\mathsf{var}(\hat{\theta}^{[2]})$ estimated using bootstrap	(0.059)	(0.090)	(0.159)
$\mathsf{var}(\hat{\theta}^{[3]})$ estimated using bootstrap	(0.052)	(0.079)	(0.146)

  As the referee points out, having accounted for the variability in our estimates of $\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$, we would expect a resulting *increase* in the estimated variability of return levels, not a decrease as suggested above. Actually, having checked this (many times!), we realised before the second submission that our estimated standard errors in the first submission (0.091, 0.127 and 0.201) were incorrect. Specifically, we neglected to incorporate our estimates of θ into the process of obtaining standard errors for $\hat{z}_r$ at all. For example, the components of ∇z_r in the delta method should be evaluated at $\hat{\lambda}_u$, $\hat{\sigma}$, $\hat{\xi}$ *and* $\hat{\theta}$: at this point, we used

$$z_r = u + \frac{\sigma}{\xi}\left[(\lambda_u r n_y)^{\xi} - 1\right]$$

  to obtain the components of ∇z_r instead of Equation (5) in the current submission, and our variance–covariance matrix for the model parameters was a 3×3 matrix for $(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi})$

only. Evaluating ∇z_r correctly, and using a 4×4 variance–covariance matrix for $(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi}, \hat{\theta})$ (where $\hat{\theta}$ is replaced with $\hat{\theta}^{[2]}$ or $\hat{\theta}^{[3]}$ and where the variance of these estimates is assumed zero), the standard errors for our return levels in the first submission *should* have been:

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
$\hat{\theta}^{[2]}$ constant	(0.057)	(0.089)	(0.158)
$\hat{\theta}^{[3]}$ constant	(0.047)	(0.077)	(0.145)

Comparing these to the standard errors which incorporate variability in our estimates of θ (first table), we can see that the estimated standard errors for the return levels have increased, as we would expect.

(ii) Standard errors for estimated return levels when $\hat{\theta}^{[5]}$) is used changed because of the error in the original bootstrapping scheme as pointed out be referee 2.

(iii) The referee comments that some of the other standard errors changed between the first and second submissions, and he/she does not know why. Again, we apologise for not explaining these changes in our report which accompanied the second submission. On re–calculating all of the values in our tables in preparation for the second submission, we realised that there were other mistakes in the calculation of standard errors in the first submission: for example, in places we had used $\mathsf{se}(\hat{\lambda}_u)$ instead of $\mathsf{var}(\hat{\lambda}_u)$ in the delta method.

(iv) The referee points out that although we corrected Equation (13) in the first submission with Equation (8) in the second submission, the standard errors in the first row of Table 2 on page 20 did not appear to change. We apologise for this – we obviously forgot to update this table completely! The correct standard errors have now been inserted.

We have checked all of this again (and again (!), with some scrutiny) and we are now confident that the standard errors in Tables 1 and 2 are correct. We apologise to both referees for our sloppiness here!

[Our response to referee 2's report is overleaf]

Referee 2's report

The Associate Editor does not think it is essential for us to address the first two points of Referee 2's report (using Süveges' estimator of θ and checking the performance of our approach under model mis–specification). Although we agree with Referee 2 that these would be interesting lines of further enquiry, to keep the paper to a reasonable length we do not consider these comments in our current work. We have been asked by the referee, and the Associate Editor, to add a conclusions section (Section 5 in the current submission) and to re–consider the type of bootstrap confidence intervals we use for z_r. Doing so has increased the length of the paper, and as we were under pressure to reduce the length of the paper after the initial submission, we leave these comments for future consideration.

As we said earlier, the main aim of our paper is to demonstrate the use of all threshold excesses in return level estimation; there are other extremal index estimators hwe have not considered (not just Süveges'), and Ferro and Segers' intervals estimator seems to perform extremely well anyway.

- The referee recommends that we use better bootstrap confidence intervals for z_r than the percentile intervals used in the last submission (Section 4.3). We thank the referee for making this suggestion; we now use the bias–corrected and accelerated (BC_a) method as proposed in Efron (1987). Details, including a practical calculation procedure for the position of the confidence bounds in the bootstrap sample, are given in Section 4.3 of the new submission (from the third paragraph on page 17).

- Minor comments

  – Page 1, line 35: "...such analyses are extremely wasteful of data". As recommended by the referee, we change this to "...analyses based on such declustering schemes..."

  – Page 1, line 44/end of page 12: We change "real–life data" to "data" or "real data", as recommended by the referee.

  – We add "Bootstrap" to the list of keywords.

  – Page 2, line 54: We replace "...can be rather unwieldy" with "...can be quite impractical", with reference to the very wide intervals that can be produced when using cluster peak excesses (as demonstrated in Davison and Smith (1990)).

  – Page 3, line 30: We discuss that our aproach, using all threshold excesses, increases estimation accuracy and precision relative to a typical POT analysis: estimation accuracy increases because of reduced bias, and estimation precision increases owing to the inclusion of more extremes (and so we get narrower confidence intervals).

  – Page 4, line 8: We change "$n \to \infty$" with "for large n", as suggested by the referee.

  – Page 4, line 43: The referee asks us to include the page number for Leadbetter *et al.* (1983). Actually, we have replaced this reference with Pickands (1975), in which the result that the GPD is the limiting distribution for excesses over thresholds was first derived.

  – We change "Coles(2001)" to "Coles (2001)" in the footnote on page 5.

  – Page 6, line 24: We replace "... the level that is exceeded once every s observations" with "...the level that is exceeded once, on average, every s observations"

  – Page 6, line 42: $2\{\ell(\hat{z}_r) - \ell(\hat{z}_0)\}$ – we remove the 'hat' from z_0.

  – Page 6, line 44: We replace "associated log–likelihood" with "associated profile log–likelihood".

– Page 7, line 28: We replace “substantial under–protection” with “substantial over– or under–protection”.

– Page 8, lines 42/45: We replace “the other $\rightarrow$ 1” and “the other $\rightarrow$ 0” with “the other approaches 1” and “the other approaches 0” (respectively), as recommended by the referee.

– Page 10, line 46: We have corrected “the the”

– Page 12, line 17: We replace “might not expect” with “would not expect”.

– Page 13, Section 3.5.6: Asymptotic independence. We choose *not* to show results for our simulation study using Gaussian AR(1) processes because of space issues. The Associate Editor has asked us to include an overall conclusion section to this work, which we agree is needed; we have also expanded Section 4.3 to include information about the bias–corrected, accelerated bootstrap confidence intervals used – referee 2 and the Associate Editor pointed out after the last submission that this Section needed attention. Since there were already concerns about the overall length of this paper, and the additional material discussed here *is* necessary, we feel that we do not have the space to expand Section 3.5.6. In fact, the plots in Figure 4 would have to be reduced in size substantially if we were to add another row of plots correspnding to the AR(1) case; we think this would compromise the ability to understand and interpret these plots.

– We have edited the caption for Table 1: it should now be clear that the units (metres) are for the estimated return levels ($\hat{z}_r$). We have *not* edited the caption for Table 2: the units (knots) are clearly for the estimated return levels as all results given in the table are for return levels, no results for the extremal index are shown here.

– The referee asks if all the digits given in the tables on page 20 are useful (we report results to 3 d.p.). We keep all results to this level of accuracy, but will be willing to reduce this if the Associate Editor thinks this is necessary.

– We have changed the size of Figure 1 so that we now have square panels, as suggested by the referee.

– The referee remarks that, in principle, it would be better to avoid words/expressions like “of course” and “obvious”; we agree and, where appropriate, we have removed such words or edited the text. NEED TO CHECK FOR OTHER “STRONG” WORDS/EXPRESSIONS!

– The referee comments that it would be good if we could spell out our ‘best practice’ explicitly in a concluding section. In fact, the Associate Editor agrees (see Page 1 of this report), and so we now include Section 5, which (we hope) pulls together the main findings of our paper succinctly, detailing what we believe is best practice as far as estimating return levels for environmental extremes is concerned.