We are very grateful to Anonymous Referee #1 for their comments and suggestions on our manuscript. We have prepared the following responses to the points raised and indicate our proposed changes to the manuscript.

Anonymous Referee #1

Received and published: 25 September 2017

General comments

The aim of this paper is to improve the ability of mechanistic rainfall models extremes to reproduce rainfall extremes. This is achieved by fitting these models to the amounts by which rainfall totals exceed a certain threshold level, or censor. Totals below this level are censored: their value is taken to be zero when the models are fitted. Applying a threshold to data is a standard approach in extreme value modelling. It's use here is a novel and interesting idea. If modelling rainfall extremes is the primary goal then using the censor to reduce the influence of small totals is sensible. Discussion paper Of course, a key issue is the choice of censor: a low censor may not achieve the desired objective but as the censor is raised the precision of estimation reduces and, in the current context, may exacerbate the parameter identifiability problems to which the fitting of these models are prone. This is analogous to the choice of threshold in an extreme value analysis and therefore it is unsurprising that in Figure 13 they consider informing this choice using a graphical approach that is common in extreme value Interactive modelling. It may be productive to explore other methods proposed in the extreme comment value threshold selection literature.

Overall I am positive about this paper. My main criticism is that better reproduction of rainfall extremes is achieved by tuning various things: censor; model parameterisation; fitting properties; perhaps even the model itself, in order to achieve this objective. There is further work to be done to provide methodology to make these choices.

Specific comments

page 11, line 288. This isn't quite correct. These weights are not optimal. However, in practice they are close to being optimal and are easier to estimate than the weights that are optimal. On that note: how are these weights estimated?

The estimation of weights follows the theory and procedure set out in the MOMFIT manual (Chandler et al. 2010). First, the covariance matrix of mean summary statistics Σ is estimated using Equation 1.

$$\hat{\Sigma} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (T_i - \bar{T})(T_i - \bar{T})'$$

Equation 1

T_i is the vector of summary statistics for the ith period. $\bar{T}$ is the vector of mean summary statistics and n is the total number of periods which in practice is the length of the dataset in years. The optimal choice of weights W would be closely approximated by the inverse of the covariance matrix given in Equation 2.

$$W = \Sigma^{-1}$$

Equation 2

However, as reported by Chandler et al. (2010), this results in highly inaccurate standard errors because the unique elements of Σ are estimated separately by Equation 1. To overcome this, a reasonable approximation for the weights is to take the inverse of the observed variance where t_i is the vector of diagonal elements of Σ.

$$\omega_i = 1/var(t_i)$$

We can update the manuscript to indicate that the estimation of weights is not optimal, but provides a robust estimate.

page 12, lines 305-308. This isn’t quite correct. The sampling distribution of the GMM estimators is approximated by a MVN distribution, i.e. there is an approximation involved and the result is for the rule that is used to calculate the estimates, rather than the estimates themselves. Line 306: Hessian of what? Lines 307-308. This sentence isn’t clear. Presumably the point is that the calculation of confidence intervals can fail in some cases. Perhaps it would be sufficient to reverse the ordering of the sentence to make the causation clearer.

Line 305: We can amend the text to highlight that the distribution of the estimator resulting from the minimisation routine is approximately multivariate normal.

Line 306: The covariance matrix is estimated with the Hessian of the least squared objective function S given on line 280. We can revise the text to make this clearer.

Lines 307-308: Indeed, the sentence was provided to highlight that the calculation of confidence intervals can fail. We will reverse the sentence as below.

On occasions that the model parameters are poorly identified, it may not be possible to calculate the Hessian of the objective function preventing the estimation of parameter uncertainty.

page 13, line 345. If you are interested in the 1000 year return level why not simulate 100 realisations of 1000 years duration?

Extreme value estimation up to the 1000-year return level is provided to indicate the potential magnitude of rarer events. Estimation up to the 1000-year return level is performed for the optimum parameter set, therefore there is no sampling from the MVN distribution of model parameters. To ensure these estimates are reasonably stable up to 1000 years, simulations have been extended to 10k years. Furthermore, simulation bands for 100 simulations of 1000 years would differ from the 100 year simulation bands shown and make the already busy figures more crowded.

For validation that the observed data at each site are sampled from the sampling distribution, it makes sense to derive simulation bands for a length of data that is of the same order of magnitude as that of the observations. Therefore, the duration of 100 years was chosen to cover the range of the data at both sites and to enable comparison between sites.

pages 14-16, Figures 4-6. There seems to be a slight upward curvature in the lines based on the 10000 year simulation. In the context of an extreme value analysis this is consistent with the shape parameter discussed on page 4 (line 111) being positive. This might be worth a brief comment. Is there any work that examines how the extreme value properties of this type of model relate to the model parameters and therefore provide a link to the theoretical basis that underpins extrapolation from extreme value models?

This is a good suggestion and we will comment on this as suggested in the discussion. That said, we are not aware of any such work, but we will briefly review the literature again to investigate.

page 17, lines 385-386. I disagree. I suppose that it depends what you mean by “poorly identified”. However, it is to be expected that as the censor is increased un- certainty about model parameters increases. If we think that we need a larger censor, because otherwise there is systematic underestimation of extreme rainfall totals, then we need to accept higher levels of parameter uncertainty.

This is a good point, although we have observed the same with uncensored models. Overall, our choice of summary statistics gives rise to apparently very well identified model parameters at sub-hourly scales. The comments here relate specifically to estimation at the hourly scale.

Our statement that the parameters are poorly identified may be overly strong. We can change this to state that confidence in the estimation is reducing.

page 18, Section 6.2. Is this level of tinkering with the choice of censor justified? Having a different censor for different levels of data aggregation feels like cherry-picking. Also, in the previous section an argument was made against a censor of 0.6mm for Atherstone but now it is being used.

The model is fitted separately for each temporal resolution which explains why we have different censors. Because of the effect of aggregation, we cannot use a model censored at one resolution to estimate rainfall extremes at a coarser one. Therefore, censored model parameters are scale dependent which is explained in section 3. Furthermore, we believe that the use of different censors at different levels of aggregation is justified on the basis that the distribution of rainfall amounts differs between aggregation levels.

The research as presented is exploratory and intended to investigate the potential for censoring in estimating extremes. In this context, a range of censors have been investigated and shown to be effective at improving the extreme value estimation of the models, and a choice had to be made for validation. Considering that, we agree that it would have been more consistent to select 0.2 mm for validation of the Atherstone hourly censored model and we will make that change.

page 19, Figure 8 caption (and elsewhere). “optimal censors” seems like a bold claim given the difficultly of choosing the censor.

We agree and will refer instead to selected censors.

page 28, line 516. The independence criterion isn’t a requirement in extreme value modelling. See Fawcett and Walshaw (2012) Estimating return levels from serially dependent extremes. Environmetrics 23(3), 272-283.

Thank you for the reference, we were unaware of this research and appreciate its relevance in the context of ours.

We note that the methods set out in Fawcett and Walshaw (2012) requires the estimation of the extremal index which appears to be subjective and potentially non-trivial. Therefore, we feel that our comparison with the standard peaks over threshold approach in which independent cluster peaks are identified is still valid. We will highlight this in the discussion with reference to this research.

Technical corrections

page 3, line 81. At this point, or perhaps even in the abstract, it is worth explaining briefly the nature of the censoring. At the moment we need to wait until page 7 for this.

We will make the following change to the text.

To test our hypothesis, a simple approach is proposed in which low observations for fine–scale data are censored from the models in calibration. For a given temporal resolution, a censor amount is set. Rainfall below the censor is set to zero and rainfall over the censor is reduced by the censor amount.

page 4, line 118. “behavioural parameterizations”. Given that you use this term later it would be worth explaining (somewhere) what this means in the context of the current paper.

This was also highlighted by Anonymous Referee #2 therefore the response below is the same as that provided to referee #2.

We have used the term “behavioural parameters” by analogy with Beven and Binley (1992). We have used the term to refer to well identified models. We have found that for well identified parameters with narrow 95% confidence intervals, simulation bands on the extreme value estimates are correspondingly narrow. As the parameters become less well identified, their 95% confidence intervals increase giving rise to extreme value estimates which deviate significantly from the observations, which in turn results in significant deviation of the simulation band upper limit. This effect is shown in Fig.11 resulting from the very large parameter uncertainty shown in Fig.12.

We will remove the reference to “behavioural parameterizations” in the context of this research and change all references to well identified parameters.

page 6, line 160. “Lower variability” may be better than “less variance”.

Agreed. We will make this change in the manuscript.

page 10, line 262. Presumably the reason for the missing data in 1974-75 was political rather than environmental. It might be worth noting that the fact that the data are missing is not expected to be informative about rainfall totals.

We do not know why the data are missing, but we can certainly highlight that this is not expected to affect the results. We will make the following change to the text.

Atherstone is a tipping bucket rain gauge (TBR) operated and maintained by the Environment Agency of England. The record duration is 48 years from 1967 to 2015, with one notable period of missing data from January 1974 to March 1975. The reason for the missing data is unknown, although it is not expected to affect model fitting and the estimation of extremes.

page 12, line 313. “ldots extreme values continued to be underestimated . . . ” might be better.

We will change this sentence to the following.

While good model fits were obtained for some low censors, extreme value estimation continued to be underestimated.

pages 14-16, Figures 4-6. The plots would be clearer if the scale on the lower horizontal axis was return level in years. The AEP on the upper horizontal axis would then be unnecessary. The scale of the Gumbel reduced variate adds no information in itself. These plots are quite crowded and

We agree that the Gumbel reduced variate adds no additional information given that the AEP is provided on the secondary x axis. We also agree that removing this and moving the AEP to the primary x axis will simplify the plots. However, we propose to keep the content of the plots unchanged as they show the convergence in estimation for all AEPs up to 0.001 for increasing censors.

page 17, line 358. I'm not sure that I would use "confidence intervals" here. Perhaps "simulation bands"? ... and say explicitly what this means, i.e. how the lines in the plot are calculated.

This was also highlighted by Anonymous Referee #2 therefore the response below is the same as that provided to referee #2. We agree with the suggestion and will change all occurrences in the manuscript.

There are in fact two issues here: if we were doing very long simulations with practically no random noise (so that another simulation would yield practically the same result), then we would have identified approximate confidence intervals. But with the shorter simulation length, both parameter uncertainty and the randomness of the model are combined in the spread we observe in the simulated statistics, so that 'simulation bands' is indeed a better descriptor.

page 17, line 379. I'm not sure what this sentence means. Are we supposed to be looking at Figure 7 for evidence of this?

The reader should be looking at Fig.6 for evidence of this divergence in estimation. This is explained with the aid of Fig.7. We will make the following change to the text.

The mean of the MVN realisations for the BL1M model at Atherstone with the 0.6 and 0.8 mm censors (see Fig.6) diverges from the optimum because of the generation of unrealistic extremes. This divergence is also observable in the larger spread of 95% simulation intervals over 100 realisations.

pages 20-21, Sections 6.2.1, 6.2.2 and 6.2.3. I don't see the point of including these sections. Section 6.2.1 shows exactly what we expect: by excluding properties that are difficult to reproduce we are able to reproduce well the properties that are not excluded. The comparison in Section 6.2.2 is unsatisfactory because we cannot compare like with like, owing to the truncation of the data but not the model. Section 6.2.3 just shows that there are clear local minima in the objective function but we can't expect to search too far in the search for confidence limits.

It is not always the case that the fitted parameters well reproduce the summary statistics used in fitting. The purpose of these checks is to ensure they do given that the data are censored. That said, given that the ability of the models to reproduce the summary statistics used in fitting at both sites is equally good, we could reduce this section by only showing plots for one site. We could then state that comparable performance is achieved at both sites.

We take the point about checking skewness. Given that we've already highlighted that skewness is not expected to be well reproduced because of the truncation of the data we are happy to remove these plots. However, we feel there is still validity in checking the proportion of dry periods as this property is strongly affected by removing low observation.

We would be happy to remove the profile objective function plots given that there is other evidence of good parameter identifiability with high confidence on the parameter estimates.

page 27, Figure 14. I don't think that these figures add much to the statistics concerning the proportions of totals lying below the censors, with the possible exception of the visualisation of the resolution of the Atherstone data.

These figures were included to demonstrate graphically how much data is removed by censoring. Given that the methodology for selecting a censor presented in this research is based on a graphical approach, these figures are useful to understand the rainfall quantiles which have given rise to well parameterised censored models. Until an alternative method is developed to optimise the censor, we feel these plots will aid other practitioners in estimating rainfall extremes with censoring. Therefore, we propose to keep these plots in the manuscript.

page 28, line 514. The rule to try to create independent peaks needs to be given earlier: before the concept appears in Table 3.

Noted. We will bring this forward so that it is highlighted before Table 3.

pages 28-30. Do we need both "Further discussion" and "Conclusions"?

We can look at combining these into one section possibly called *Further discussion and conclusions* or just *Conclusions*.

page 31, line 591. Is the first inequality sign the wrong way round?

Yes, it is. Thank you for highlighting this. We will correct this in the manuscript.

page 32, Figure A.2. Below and to the right of the plot is says that 1/L has an exponential distribution, which, according to the description of the models on page 8, isn't true.

Thank you for highlighting this inconsistency. We will correct this and update Fig.A.2. We also notice that the parameterizations for X and L are listed wrongly in line 225. We will revise this as follows.

Both X and L are assumed to be independent of each other and follow exponential distributions with parameters $1/\mu_x$ *and* η *respectively.*

References

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6, 279-298, 1992.

Chandler, R., Lourmas, G. and Jesus, J.: MOMFIT Software for moment-based fitting of single-site stochastic rainfall model fitting, User guide, Department of Statistical Science, University College London, London, 2010.

Fawcett, L. and Walshaw, D.: Estimating return levels from serially dependent extremes, Environmetrics, 23, 272-283, 2012.