The manuscript presents an approach based on entropy for choosing the most suitable statistical models to represent partial duration series of streamflow. In particular, it proposes to evaluate the combined entropy of the statistical models describing the arrival of peaks above a certain threshold and the magnitudes above this threshold.

The idea is interesting, especially because it advocates using additional criteria (i.e., the capability to represent occurrences of events exceeding a threshold) to the goodness of fit of theoretical distribution calibrated to magnitude exceedances above the threshold. However, the study has some issues which prevent reaching substantial conclusions. They are described below.

- Calculating entropy constitutes an additional step to the usually applied procedure in this field. The value of performing this additional step should be made clear. As I stated above, I see value in the evaluation of an additional criterion to the goodness of fit of theoretical distribution calibrated to magnitude exceedances above the threshold. However, what advantage does it actually bring with it? Does this method for choosing the most suitable statistical model improve its predictive power? The authors claim it does, but the support of this claim is not clear to me (see the next comment).

In the present study, the authors have proposed entropy as alternate goodness of fit (GOF) measure for threshold identification in partial duration sampling of flood frequency analysis. The prerequisites for threshold identification in PDS, such as satisfying the graphical tests proposed by Lang et al. (1999) and independence of peaks by several guidelines or other statistical tests, remain the same. The next step of evaluating the degree of fitness of different distributions to model the magnitude of exceedances generally involves the application of several GOF measures such as Chi-squared, Kolmogorov-Smirnov (KS) and Filliben Correlation (FCC), Anderson-Darling (AD), modified AD, root mean square error (RMSE), mean absolute error (MAE), relative mean absolute error (RMAE), different information criterion, etc. The present study recommends maximizing the total entropy of both the models of the PDS instead of calculating different combinations of error criteria as suggested in several pieces of literature. So the calculation of entropy has been proposed as an alternate step in threshold identification of PDS in flood frequency analysis. To this end, the authors have compared results attained from the proposed methodology with those obtained applying combinations of several error statistics (Table 3 and Figure 7 of the manuscript) for the study area considered in this work. LP 3 has the maximum total entropy at most thresholds, where P3, GP, and GEV possess the second, third, and fourth, respectively (Figure 6f). Similar results are obtained from the statistical ranking of distributions also (Figure 7). Comparable results are observed for individual distributions, such as LP 3 has the maximum total entropy at a threshold of 710 m3/s. It best described the exceedances extracted at 700 and 710 m^3/s as per other test statistic's rankings.

The study involves the evaluation of an additional criterion to the goodness of fit of theoretical distribution calibrated to magnitude exceedances above the threshold. It's observed that the threshold with the maximum entropy of Model 2 (i.e., used for the magnitude of exceedances) is different from the one at which the total entropy of both the models is the maximum (Figure 6). The latter, selected as the optimum threshold, is close to other values obtained from existing literature. So combining entropy of Model 1 (arrival rate of peaks) improves the accuracy of threshold identification. For example, the entropy of Model 2 is the maximum at a threshold of 900 m^3/s for LP 3 distribution, while combining entropy of both the models maximizes entropy at 710m^3/s, which is similar to the optimum thresholds found from the existing guidelines and studies carried out in the study area previously. Also, the optimum threshold obtained from the proposed method lies in the stabilized region of Figure 9, which justifies the predictive ability of the model.

- The authors claim to discuss the predictive ability of the statistical model selected by means of the proposed entropy-based approach in Figure 9. The figure shows the flow value associated to 50 and 100 years return period, calculated by means of a generalized Pareto distribution calibrated to exceedances above a set of different thresholds. Confidence intervals of the estimates are also displayed. I do not understand what this figure tells about predictive ability. I would be happy to hear about it; in case I am missing something obvious. First of all, Log Pearson 3 is the most suitable statistical distribution according to the values of entropy, whereas results for generalized Pareto are shown here. Then, where do we see in Figure 9a better predictive performance of the distribution suggested by the entropy metric? Also, its predictive performance is better compared to what? I guess it should be better compared to the performance of the statistical model that would have been chosen based on goodness-of-fit metrics displayed in Figure 7 (see the next comment about the interpretation of those results).

  In bootstrapping, we repeatedly sample from the observed dataset, with replacement, forming a large number (B=1000 in this study) of bootstrap datasets, each of the same sizes as the original data. The idea is that the actual observed data takes the place of the population of interest, and the bootstrap samples represent samples from that population. To use bootstrapping for analyzing the predictive ability of an estimate, we fit our model to the original data and fit the model to each of the B bootstrap sample datasets. We calculate values for 95% confidence intervals (CI) and plot the confidence interval from the estimates obtained from the B samples. The predictive ability of an estimate can be checked if it lies within the upper and lower limit of the CI. In the present study, similar bootstrap analysis is carried out (details are described in the following comment) and 95% CI for 50 and 100-year period quantiles, and the actual values are calculated at various thresholds and plotted as shown in Figure 9. The authors have also included the 95% CI of 50 and 100 year return period quantiles for LP 3 distribution in the revised manuscript. In this plot, the threshold obtained from the proposed entropy approach is compared to the performance of the statistical model fitted at the optimum threshold that would have been chosen based on goodness-of-fit metrics (Figure 7).

- I also do not understand how the bootstrapping was performed: could you provide a number for the length of data used for each resampling (line 344)?

  The non-parametric bootstrap sampling procedure is applied in the study;
  - N_b bootstrapped series (N_b = 1000) of X_p peaks are obtained by bootstrapping (i.e., resampling with replacement) of X_p original peaks derived at each threshold level. So bootstrapped series are given by $\{X_p\}_j$ with j = 1, 2, 3…, N_b, and p is the number of flood exceedances at each threshold level; i.e., each bootstrapped series has an equal number of peaks as that of the original sample.
  - For each bootstrapped series $\{X_p\}_j$, distribution parameters, and 50, 100 -year flood quantiles are estimated.
  - The values of the estimates for a 95% confidence interval are calculated and plotted.

  These description has been included in the revised manuscript.

- In addition, the authors state at line 365 that the proposed method “gives more accurate optimum threshold values”. Based on what facts do they claim the threshold identified from the entropy metric to be more accurate? What is their reference value?

  Considering the entropy of model 1, i.e., the arrival of peaks instead of taking only the entropy of distributions used for modeling exceedances gives more accurate optimum threshold values. The study involves the evaluation of an additional criterion to the goodness of fit of theoretical distribution calibrated to magnitude exceedances above the

threshold. It's observed that the threshold with the maximum entropy of Model 2 (i.e., used for the magnitude of exceedances) is different from the one at which the total entropy of both the models is the maximum (Figure 6). The latter one, selected as the optimum threshold, is close to other values obtained from some existing pieces of literature.

- Lines 358-363 simply discuss thresholds identified by means of alternative methods. If the value from the operational guidelines of Lang et al. (1999) (i.e., 730 m^3/s, line 359) is used as reference (although this is also just another method) then Langbein (1949) would still provide a more accurate threshold (716 m^3/s) than the proposed method (710 m^3/s). Please clarify.

Here, the authors compare only the threshold values obtained from other existing guidelines or previous studies. They have compared only the values of optimum threshold obtained from different methods are close to the one proposed in the study. Like for LP 3 distribution, model 2 has the maximum entropy at a threshold of 900 m^3/s, while combing model 1 with this, the total entropy is maximized at a threshold of 710m^3/s, which is close to the values obtained from some existing methods. So considering the entropy of model 1 improves the accuracy of threshold identification for the proposed method. A similar explanation is added in detail in the revised manuscript.

Additional points

- The proposed approach involves several steps which rely on visual observations and graphical analyses. These usually imply a high degree of subjectivity and difficulties to apply them to large datasets. It occurred to me that the approach described in section 2.4 to identify independent peaks is the same adopted by recent papers which leverage the Metastatistical Extreme Value framework to estimate flood magnitude and frequency from the whole series of ordinary peaks (i.e., with no need to define a threshold). Given that this novel statistical approach is gaining momentum, and that differently from the approach proposed here it can be completely automatized, it may be good to spend some words to justify the importance of identifying partial duration series by means of the classical peak over threshold methods.

The proposed approach involves identifying an initial threshold range from the operational guidelines proposed by (Lang et al., 1999), which involves visual observation of mean residual life plot (MRLP). Domain 3 of Figure 3 can be identified once we extract independent peaks at each threshold level starting from the minimum daily discharge and calculate the length of the PDS. A visual inspection is needed for the MRLP to identify a possible range of thresholds where the optimum threshold might lie. Before the mere visual inspection, the numerical results obtained for plotting both the graphs should also be analyzed.

The application of the metastatistical framework is adopted by some recent researchers to estimate flood magnitude and frequency from the whole series of ordinary peaks. However, the classical peak over threshold method is still popular among researchers while modeling with discharge data samples, and it involves a few uncertainties that need more attention. So in this work, the classical PDS sampling approach is further explored applying the concept of entropy. The application of this entropy-based concept can also be examined in the Metastatistical Extreme Value framework, which is out of the scope of this technical note. However, the author would like to address it in future research work along with the automation of the proposed entropy-based method.

Some suggestions concerning the structure of the paper:

- More precise explanations of what is shown in the figures and how it enables to reach the stated results are needed. Just to give two examples: line 240: how did Kendall's Tau verified the independence of the series? How do we see it? line 286: why finally the Poisson and not the Binomial distribution is chosen for the arrival of events above a threshold?

The detailed procedure of Kendall's Tau test and Poisson hypothesis test is included in the revised manuscript. PDS at those thresholds where Kendall's tau is less than the critical value at 5% significance level are considered to satisfy the independence criteria. Similarly, based on the upper and lower limit of the dispersion index value, Poisson or Binomial distribution is selected. Details of these are described in the revised manuscript.

Figures 1 to 5 display results of standard procedures which could be easily summarized with a few words in the text. Although this is a Technical Note where technical details shall be provided, Figure 2b-d only shows examples of results for arbitrarily chosen thresholds and Figure 4b is simply a zoom of Figure 4a. These figures could be deleted, which would help highlighting the actual results of the approach proposed in the paper (Figure 6).

The authors have incorporated such modifications of figures in the main manuscript.

- Figures with several panels could be condensed. For example, Figure 6 could be condensed to Figure 6f only; Figure 7 can be condensed in one single panel displaying total rank only.

The authors acknowledge this suggestion, and Figure 6 is modified in the revised manuscript. In Figure 7, the ranking of distributions at four different thresholds is shown for illustrative purposes. It's been changed in a single panel in the revised manuscript.

- Several minor issues exist in the paper, especially related to correct and precise use of language (e.g., 22, 68, 163, 167, 174, 128-129, 130, 212, 216, 255), definition of symbols and units (symbols shall be introduced the first time a variable is named, e.g., t is only defined at line 274 although appearing in Figure 1), differences between statements on the same subject (e.g., lines 88 and 314), motivations for showing these specific plots, given that many are examples for, e.g, different threshold (e.g., Figure 2 and 7). I do not detail them all here given the prior need to address the major issues described above. A carefully revision of the manuscript is however recommended.

The authors would like to thank the reviewer for suggesting these corrections. Based on this, the manuscript is thoroughly revised to address all these minor issues related to language, symbols, units, etc.

References

Lang, M., Ouarda, T. B. M. J. and Bobe´e, B: Towards operational guidelines for over-threshold modeling, J, Hydrol, 225, 103–117, 1999.