Response to Reviewer 1:

Thank you for your comments that has helped us to improve the manuscript. We hope that the following changes that we have made, make the manuscript now easier to read and show the novelty we are bringing more clearly. The responses to your comments are shown below in blue font.

1. Overall, the writing is mediocre and needs improvements. This is not only about English but the way the story is told.

   We have made several changes in the manuscript that make the flow easier to follow. We hope that our main goals and conclusions are now clearer to the reader.

2. Abstract needs to be revised to better discuss the methods and major results. Some quantitative results (numbers) are needed; as of now, everything has been discussed subjectively.

   The abstract has also been updated accordingly, by discussion the methods and the results. Also, numbers are given to discuss the finding more objectively.

3. In the Introduction section, it is not clear which approaches and methods you are addressing in different references. Please clarify. My comments in the PDF can clarify what I specifically refer to.

   Thank you for your detailed comments. We have tackled them in the new version of the manuscript. Some relevant responses from the pdf comments are listed below:

   Line 36: Moderate rain on snow also makes steep risings, but it is not related to the storage. Please clarify this with appropriate references.

   We have added a short explanation in the introduction about moderate rain on snow that can cause steep rising.

   Line 62: Catchment land cover, season (time) as well as the rainfall type affect this ratio. How did you separate them?

   Depending on the combination of these factor; catchment land cover, rainfall type, soil type and initial conditions etc., the catchment will respond differently and show a distinctive flood hydrograph. We are calculating the p/V ratio based on the mean daily flows (MDF) for each event recognized in the monthly instantaneous peak, so their effects are already reflected (and included) in the p/V ratio.

   At line 68: Why MHQ?

   We include the mean maximum flow because it is easier to find the best linear model setting that estimates adequately all flood quantiles. Since we have a variety of predictors (not all are included in this paper), it is easier to find the best setting based on the MHQ that will represent both low and high flood quantiles. If our aim was three quantiles, then the selection of the predictors may change for each case. In case one quantile is used as a target, then most probably the other quantiles will be under/over – estimated. Another important point to keep in mind is that other studies comparing IPF estimation method also analyse the annual maximum flow on average, and do not compare the quantiles. Thus, for a fairer comparison, the MHQ is used as a target variable.

Line 90: Why? Sometimes IPF can be bigger than the annual maximum peak even if it did not happen in the annual maximum peak day.

Yes indeed. However, correcting each event as in application 1, can lead to very high correction values for the MDF if smaller events are not properly distinguished. Thus, it might be more robust to correct the annual maximum daily peak, even though they may not occur at the same day at the IPF. That is why we compare these two applications with one another, to see which of them performs better in our catchments.

Line 209: Is there any snowmelt in May and June from high elevation areas?

Line 220: Snowmelt in the winter or spring?

The main assumption is that the snowmelt occurs in winter (November to April). Of course, there might still be some snowmelt in May and June in the alpine catchments, and these events are classified as summer events. However, this affects only a very small portion of our catchment. We have discussed this shortly in the updated manuscript.

Line 228-230: But it is not only related to the size. What does the increasing annual maxima mean? Annual maxima is only one value per year

We mean here that mostly in small catchments, the annual maxima are wrongly chosen in winter season by the mean daily flows, rather than in summer season as observed in the instantaneous peak flows.

Line 402: In the methodology section, this was mentioned as 1,000 times. Please check and be consistent

Yes, in the methodology section we mean 1,000 realisations, and 1,000 realisations are used for the uncertainty analysis. However as explained in Line 401, this is just an example with a reduced number of 100 realisations to explain the uncertainty sources visually.

4. Please explicitly state the objective and research gap in the Introduction section.

   The objective and the research gap in the introduction section are better explained.

5. In addition to drainage area and topography, what other factors affect the peak/mean ratio? Why do you solely study the two factors?

Prior to our study we have included a total of 56 descriptors (climate, soil, topography and hydro-geological descriptors) that were investigated in a stepwise regression to determine their importance on the peak ratio. As it turns out, area and elevation proved to be more important and that's why they are our only focus presented here.

We have included a short explanation in the updated version of the manuscript to make it clear why we have chosen these descriptors.

6. Instead of using the three return periods, why not using the historical events for the analyses between flows and drainage area-elevation?

   The focus of the paper is the flood frequency analysis with the aim to improve the extreme flood events for design purposes. That is why the focus is mainly on discharge levels with different return periods.

7. Methodology section lacks proper references for the assumptions, methods and equations.

   We have included more explanation in the methodology section.

8. There is a mention of the availability of IPF monthly maximum flows, but monthly is a too large timescale for instantaneous peak flow analyses. What is your justification for using that?

   I think there is a misunderstanding here. IPF data are provided as non-equidistant monthly instantaneous peak flows, that means for each month we have the maximum recorded peak flow (in m^3/s). This means for each year we have 12 maxima instantaneous peaks, and we can select the total maximum for the annual maximum series or the maxima for the summer and winter seasons, each. We have stated this clearly in the updated version of the manuscript.

9. How were IPF_{stat} and MDF_{stat} Please clarify.

   We have clarified this in the updated version of the manuscript. The *stat* refers shortly for statistics. Statistics on annual maximum series derived from both IDF and MDF series are used for the correction: like for instance the mean maximum discharge (MHQ), or the L-moments and so on.

10. Equations 2-3 need references. Have they been developed by the authors or others (need references)?

    Equation 3 are developed by the authors, while Equation 2 is motivated by the results of Fischer et al. (2016) and Fischer (2018). We have clarified this in the updated version of the manuscript.

11. Any rational for using the GEV distribution?

    According to Maidment 1993, three distributions are reasonable choices for describing flood flows namely: Generalized extreme value (GEV), lognormal and Pearson type 3 distributions. The previous study performed by Ding et al. 2015 in

Lower-Saxony Germany, showed that the GEV had highest p-value (and acceptance of the distributions) in comparison to the other two types. Moreover, other studies like Villarini et al. (2011) or Haktanuer and Horlacher (1993) have also used GEV in several catchments in Germany. In our catchments we have performed the Cramer-von-Mises test for both original and corrected series and GEV was accepted in all of them (p-value = 5%). Since, we have relatively long observations (up to 148 years), we have employed GEV due to its flexibility to be fit different tails, as the shape parameter is adjusted independently for each gauge. A short explanation on why we have chosen the GEV is included in the updated version of the manuscript.

Lastly, we would like to point out that we choose the p/V-Lmoment as the best method (correcting the L-moments of the mean daily flows) in order to have a more universal method which is not affected by the assumption of a probability distribution.

12. Why did you use a linear regression model in Equation 3?

    Typically, also in the literature, the ratio between the IPF and MDF statistics (also called the peak ratio) is modelled by a linear model based on different catchment or climate characteristics. This has proven to be successfully for many applications. On the other hand, we have applied these models, because as in in Figure 3-4 there appears to be a high correlation between the error IPF-MDF and the logarithm of the catchment area.

13. What is Q_{suc} in Equation 4?

    Equation 4 is the model proposed by Chen et al. (2017) and is our reference for improvement. Their methodology (also called the slope method) corrects the IPF series only based of the MDF information. They calculated a slope by using the mean daily peak discharge (Q_{peak}), mean daily discharge one day before the peak (Q_{pre}) and the mean daily discharge one day after the peak (Q_{suc}). We have stated this clearly in the updated version of the manuscript.

14. In line 92, what does “IPF and MDF do not necessarily overlap” mean? Do you mean in terms of their timing or magnitude? Please clarify.

    One of the main assumptions for estimating IPF base on MDF series, is that the peaks of both data occur at the same dates. With the expression “IPF and MDF do not necessarily overlap” we mean exactly this, that the peaks of the two series do not necessarily happen on the same day. We have clarified this in the updated version of the manuscript.

15. Line 143: What does “annual maxima from monthly maximum” mean?

    As mentioned before the IPF series are maximum instantaneous peak given for each month. To compute the annual maxima, we compute the maximum from 12 values (one maximum instantaneous peak for each month) for each year- Like this we can built the annual maximum series (for the flood frequency analysis).

On the other hand, for the MDF series are daily average flow data. Here we select for each year the maximum average daily discharge observed and so we obtain the annual maxima series.

16. Line 148: To what extent, does the sample size change the uncertainty? Is 1,000 a sufficient sample size?

    When comparing methods in terms of their uncertainty, the number for random resampling will influence all methods similarly. When comparing sources of uncertainty, 1,000 realisations are enough to shed light if the linear model has higher uncertainty than the sample/parameter uncertainty. For the overall uncertainty, there were in total 1,000,000 realisations. Of course, the uncertainty ranges might change slightly if more realisations are included, but previous test that we have conducted have shown that 1,000,000 realisations are enough to capture the overall trends of the uncertainty.

17. Lines 165 and 189 are inconsistent in terms of the number of discharge stations (648 vs 653). Which one is right?

    We are sorry for the confusion; we will clarify this better in the updated version of the manuscript. The right number of discharge stations is 648.

18. Overall, there is a lack of connectivity between the subsections under Results section. This section needs a better flow.

    We have restructured a bit the section of the results in the updated manuscript and hopefully the flow within the results section is now easier to follow. The result section is now separated in two three parts: Part 1 – estimating the mean maximum flow (MHQ), Part II – estimating the GEV parameters and flood quantiles, and Part III – assessing the main source of uncertainty and the overall uncertainty range of the best selected model. For each Part I and II, we first start with an analysis on the MDF series, to see how well they match with their respective IPF (also the influence of area and elevation is discussed) and then we assess the performance of the models. In Part I we focus on the MHQ to find the most suitable predictors and linear models, in Part II we use the predictors of Part I to assess the performance of the models in terms of GEV parameters and flood quantiles. In Part II the best model is selected. Lastly in part III the uncertainty of the best model is analysed.

19. Section 4.3: It is expected that because the two databases (IPF and MDF) are different, their distribution parameters are different too. What is the main reason for comparing the parameters of distributions? A more proper comparison should be on quantiles (different return periods).

    A comparison between the quantiles is already given in Figure 10, Figure 11 and Table 6. Nevertheless, one must keep in mind that the quantiles are estimated from the fitted GEV distribution, and to understand why some quantiles are not properly represented, one should look at the GEV parameters. For instance, in Figure 10, you will see the mean error over all quantiles for MDF is negative. This is explained by Figure 8 and 9. It is clear the GEV location parameter from MDF series is underestimating the IPF-location parameter. This will affect all quantiles as the distribution is shifted to the left, hence all quantiles are underestimated.

20. Figure 13: What is the main reason for similarity among IPF-bs, MDF-bs, LM-bs-full and LM-bs-bs in different HQ years?

    The reason for this similarity is that the main source of uncertainty is the sample and parameter uncertainty (MDF-bs). LM-bs-mean, which illustrates the uncertainty only due to the linear model fitting (here the sites are resampled 1000 times in space before fitting the linear model), is considerably lower than the MDF-bs. This explains why the LM-bs-bs is very similar to MDF-bs. The LM-bs-full is just propagating the MDF-bs uncertainty through the existing model, thus the uncertainty in this case will still originate (and be similar) to the MDF-bs. On the other hand, the IPF-bs and MDF-bs are slightly different from each other, where MDF-bs is slightly less uncertain with fewer outliers than IPF-bs.

21. Figure 14: What is the difference between the median of the three HQ-years in each part of the confidence interval? They seem similar in the boxplot median.

    For each of the 486 validation sites, the median error over all realisations is computed and shown in the boxplot-median (for each flood quantile). The median error over all realisation is reflecting the same behaviour as the actual error as obtained in Figure 11. In Figure 11 the boxplots of three quantiles were also similar to one another, with clear differences in the outliers and slightly larger error spread for the higher quantiles. The same is true for the median error of all the realisations.

    For each site, apart from the median error over all realisations, the 2.5 % and 97.5% error (here referred to as the lower CI and upper CI) quantiles are as well computed. The boxplot lower CI and upper CI – show the error spread among all sites. Important to see here is as the flood quantiles are increasing, the lower CI median will get lower (so higher underestimation), while the upper CI median will get higher (so higher overestimation). This means that the higher the flood quantile, the higher the error and the uncertainty. However, we also see here that for HQ100, the median of lower CI is not symmetrically mirrored in the upper CI (as is the case for HQ10 and HQ50). This means that the errors are positively skewed.

22. Are all methods and approaches sensitive to the database type? Can those be generalized to other catchments? If so, what are some considerations?

    This is indeed an interesting question. The method proposed here is more sensitive to the flood typology of other catchments, which indirectly is mirrored in the database. In theory the predictor p/V, as it as a normalized predictor, should work well for other catchments as well. However, if the dominant floods in a catchment have a timescale less than a day (say flash floods with durations short than a day) then the daily measurements of the flow will not capture adequately the flood dynamics. Hence the linear models based on the p/V predictor may not yield good results. This was also the case in our catchments with areas lower than 100km^2. In other cases, when the flood timescale is larger than a day, then the p/V predictor should be able to capture the flood dynamics. Still, attention must be paid to the baseflow separation, to make sure that the calculated p/V predictor is representative of the catchment behaviour.

    Another thing to keep in mind, is what gauges and most important how many gauges one should group together for the fitting of the linear model. In the optimal case that the p/V predictor describes the flood dynamics correctly at each catchment, the question becomes how good one linear model can represent the whole group of catchments. Although L-moments are considered more robust than parameters or quantiles, it might be that L-moments are considerably different within the catchment group, then it makes sense to break the group down in more subgroups to better capture the L-moments. In this case we suggest the flood index clustering as suggested in Howking and Wallis (1997).

23. All acronyms and abbreviations should be spelled out in the keywords, figures, tables and headings.

    We will make sure that all acronyms and abbreviations are spelled out in the new version of the manuscript.

24. Please italicize all parameters and coefficients throughout the text.

    We have italicized all parameters and coefficient in the updated version of the manuscript.

References:

Chen, B., Krajewski, W. F., Liu, F., Fang, W. H., and Xu, Z. X.: Estimating instantaneous peak flow from mean daily flow, Hydrol Res, 48, 1474-1488, doi: 10.2166/nh.2017.200, 2017.

Ding, J., Haberlandt, U., and Dietrich, J.: Estimation of the instantaneous peak flow from maximum daily flow: a comparison of three methods, Hydrol Res, 46, 671-688, doi:10.2166/nh.2014.085, 2015.

Fischer, S.: A seasonal mixed-POT model to estimate high flood quantiles from different event types and seasons, J Appl Stat, 45, 2831-2847, doi:10.1080/02664763.2018.1441385, 2018.

Fischer, S., Schumann, A., and Schulte, M.: Characterisation of seasonal flood types according to timescales in mixed probability distributions, J Hydrol, 539, 38-56, doi:10.1016/j.jhydrol.2016.05.005, 2016.

Haktanir, T. and Horlacher, H. B.: Evaluation of various distributions for flood frequency analysis, Hydrological Sciences Journal, 38, 15–32, https://doi.org/10.1080/02626669309492637, 1993.

Hosking, J. R. M. andWallis, J. R.: Regional Frequency Analysis, Cambridge University Press, https://doi.org/10.1017/CBO9780511529443, 1997.

Maidmennt, D. R.: Handbook of Hydrology, McGraw-Hill, New York, 1993

Villarini, G., Smith, J. A., Serinaldi, F., and Ntelekos, A. A.: Analyses of seasonal and annual maximum daily discharge records for central Europe, Journal of Hydrology, 399, 299–312, https://doi.org/10.1016/j.jhydrol.2011