First of all, the authors would like to thank the reviewer for the detailed and constructive review. The reviewer mentions valid points. We (the authors) have added a response to each of them. If the reviewer agrees with our interpretation of the comments and the responses, we will solve most be making textual changes. The two main points that deviate from this are:

- Regarding the prior, we suggest doing more than just textual changes: We suggest changing the prior weakly informed prior described in Appendix A for an uninformed prior on location and scale, and an informed prior on the shape parameter. A small test shows this give similar good results, while it simplifies the method regarding the prior.
- While the suggestion of using the expert judgments as prior makes sense from a Bayesian statistics point of view, it would reduce the effect of expert judgments in this EJ-study. Therefore we choose not to do so. Hopefully this will be clarified after reading the full response.

The paper provides the result of an interesting experiment in which flood experts are asked to guess the flood frequency curves for several sites in a region without access to discharge data, and the information is then used together with observed maximum annual flood peaks to improve the "credibility" of the estimated tails of the distributions. A procedure is developed to use expert opinions on several tributaries and transform them into an estimate for a downstream gauge.

The paper is original, as far as I can tell, and deals with an important issue in flood risk assessment, i.e., the formal use of expert opinion in flood frequency analysis. Even though I liked reading the paper, there are some parts that, in my opinion, need to be improved, clarified, better explained or discussed before publication. My main concerns are:

1) I am not sure that the ability of the expert in providing a judgement on the flood frequency curve can be measured by her/his ability in guessing the 10-yr flood in absolute terms. If one wants the expert to help in reducing uncertainty in the tails of the distribution, she/he should inform us on how large floods may compare to small floods, by reasoning on the driving processes. In the end, it is the shape of the flood frequency distribution that's hard to get with local data, not the location. The proposed method seems to be tailored for getting the order of magnitude right, i.e., the flood magnitude in m^3/s, but not how surprising can large extreme events be compared to the more frequent ones.

This is a valid point, and a fundamental point to SEJ in general. We are trying to assess the expert's ability of estimating out of sample events from their ability to estimate events in the frequent/observed range. There are different approaches possible for doing this. Our choice is in line with Cooke’s philosophy, that says expert judgments should be as little perturbed as possible. This means using answers close to how they are estimating, which means it is preferable to estimate the 1000-year event directly than estimating a ratio or tail shape and deriving the 1000-year discharge (or other extremes) from that. This is further discussed in the detailed comments below.

2) the expert information is accounted for as data (part of the likelihood) using an ad-hoc procedure, which seems to me inconsistent with the Bayesian way. Why not accounting for expert judgement as prior information? That would be the natural Bayesian way to do it: since the experts give their estimates without using discharge data, this can be considered as prior information.

I agree that considering the expert judgments as priors, and the observations as data, would be a more Bayesian approach. It does however raise two difficulties in the context of this study:

1. Just as mentioned in the last point, we want to keep the expert judgment as unperturbed as possible. This means that we would need to create a prior that reproduces the 10-year and 1000-year uncertainty estimate of the experts, which would involve a complicated prior or an ad-hoc approach similar as currently described in the appendix A.
2. A two-step approach of creating an EJ-prior and updating this with observations would in some cases make the outcome insensitive to the expert judgment. This is the same issue currently described in lines 196 – 205 and appendix B. So while it would remove this subjective component from the model, it would also tip the balance towards data and reduce the influence of expert judgments in the results.

We appreciate that considering EJ as priors would be more true to the Bayesian approach, but in the context of this study, which is defined around expert judgment, we'd rather keep EJ and data together in the likelihood of Eq. 6 such that we have the option to weigh them. Note that this point is further discussed in the response to the reviewer's line 173 comment.

3) given the procedure proposed, the tail of the distribution is controlled by expert judgement with a strength that is related to the subjective choice of the weight given to the expert "data" compared to the observed data. The result of the procedure is then assessed as credible/reasonable, but how could it not be so? From what I've understood, the procedure seems to allow a way to tweak subjectively the shape of the flood frequency distribution.

The weighing factor between data and EJ is indeed a modellers choice, needed to find a balance between data and EJ when the two don't 'align' nicely. We acknowledge that this is a subjective choice, or model choice, but so are many other model choices in this study, such as using a GEV-distribution. To substantiate this, we added the sensitivity analysis in appendix B

4) the results are not assessed against a benchmark. Why not using regional flood frequency analysis as a benchmark?

The benchmark of this study is the downstream discharge at Borgharen, used in two ways:

- as historic observations (see Fig. 5)
- as design flows from GRADE (as compared in Fig. 6)

Both add value, GRADE as it gives an estimate for large return periods, the observations as they are not a model result but measured data (and therefore cannot be judged for not 'foreseeing' the July 2021 event). The discharges at Borgharen are calculated from the experts' estimates for the discharges of the river tributaries, their dependence, and a sum-factor.

We'll make it clear to the reader that GRADE is a regional flood frequency analysis that includes historic events.

5) some of the methodological steps are unclear, sometimes, and should be properly explained (see the detailed comments below).

Detailed comments:

line 8: MCMC is just a tool. I would say here that you use Bayesian inference.

Thanks, will adopt this.

line 17: the 2021 peak at Borgharen is the highest but does not seem surprisingly high, looking at Figure 5. I think the same event has been much more surprising in other, smaller catchments. Even though it is surprising for the summer season, as I understand, your analysis later is not done accounting for seasonality. I would even expect that, if asked for the summer flood frequency curve, the experts would underestimate the probability of such an event.

That is right, it is more surprising in summer (the previous summer max was about 2000 m^3/s), and as well for the smaller contribution catchments. We indeed chose to not distinguish seasonality, as it would double the number of estimates which we think would have been too much.

line 30: here the text suggests that hydrological model simulations outperform statistical methods in flood frequency analysis. Has this been demonstrated in the literature? As far as I know, statistical models tailored for flood frequency analysis are more accurate than other methods both in gauged and ungauged basins (see Bloeschl et al., 2013, ISBN:9781107028180). Besides, despite some advantages, you clearly show limitations for the hydrological modelling approach in the discussion until line 45. Since the accurate estimate of the distribution tails is of interest, why don't you mention regional flood frequency analysis and inclusion of historical events as ways of increasing the robustness (and reducing the uncertainty) of the estimates? Besides, aren't design flows available from a regional frequency analysis in the area, e.g. to be used as a benchmark?

The GRADE model is the prevailing model that provides design flows for flood defence design. That is why the study's results are compared to it in Section 4.4. The black distributions in Figure 6 named "WBI-statistics" are from GRADE. The legend label should be changed to GRADE to avoid confusion, and we'll make sure that it is clear to the reader that the benchmark is a) GRADE, as b) the historical observations at Borgharen.

Furthermore, we did not mean to suggest that hydrological model simulations generally outperform statistical methods and will reword this. Note that GRADE is a combination of a statistical method and hydrological model, which might not have added to the clarity here.

line 43: I don't get the factor 3 vs. 1.4 sentence. What is the "outcome"?

We'll specify the quantities.

lines 65-68: spoiler alert! I would move this sentence after the results section.

We will (re)move this.

line 79: I don't get the meaning of the sentence "The discharge estimates for this catchment are therefore only used for expert calibration, as the flow is part of the French Meuse flow".

In this study, we modelled the overall catchment as a number of sub-catchments that flow into a main branch. The Semois sub-catchment is part of the larger French Meuse sub-catchment, i.e., it flows into the French Meuse tributary before this tributary enters the main branch. Therefore, it’s not part of the sum-model (Eq. 1), as we would be double counting discharge. It is however a sub-catchment with a significant size and good data, which is why we did use it for expert calibration (i.e., comparing expert 10-year estimates to data).

We will clarify this or move the sentence down after were the model is introduced, where it will make more sense.

line 85: I would add a table here in the main text summarizing the data provided to the experts.

Will adopt this.

line 107: not having some more details on the construction of the correlation matrices is a pity. It would have been wise to publish that paper first.

That’s right, but the other way around would be so as well. The two are related, too big to publish together, and trying to time them together is difficult with external factors. We found this order the most logical (bigger picture first, then zooming in on the details of dependence models and elicitation).

line 109: Each variable is modelled by a marginal distribution, it is not a distribution.

Will remove the “univariate” and take “marginal” out of the brackets.

Section 3.2: I am not sure that the ability of the expert in providing a judgement on the flood frequency curve can be measured by her/his ability in guessing the 10-yr flood in absolute terms. If you want the expert to help in reducing uncertainty in the tails of the distribution, she/he should inform you on how large floods may compare to small floods, by reasoning on the driving processes. In the end, it is the shape of the flood frequency distribution that's hard to get with local data, not the location. Your ranking seems to me tailored for getting the order of magnitude right, but not how surprising can large extreme events be compared to the more frequent ones. I know this cannot be done now but I would have asked the experts to guess the ratios between the 10-yr event and the mean event, and between the 100-yr event and 10-yr event, and so on, in order to get their perception on the shape of the distribution. Maybe you could discuss the idea in the discussion section, if you see that fit.

As discussed in point 1. We will add something in the discussion on this. From a catchment hydrology point of view it would perhaps make more sense to estimate ratios, as different runoff processes become less or more important. However, from a statistical and expert judgment point of view, were we want to quantify a statistical model for extreme river discharges, it makes most sense to try and estimate the components in that model more directly.

line 151: "a training exercise"

Will be corrected.

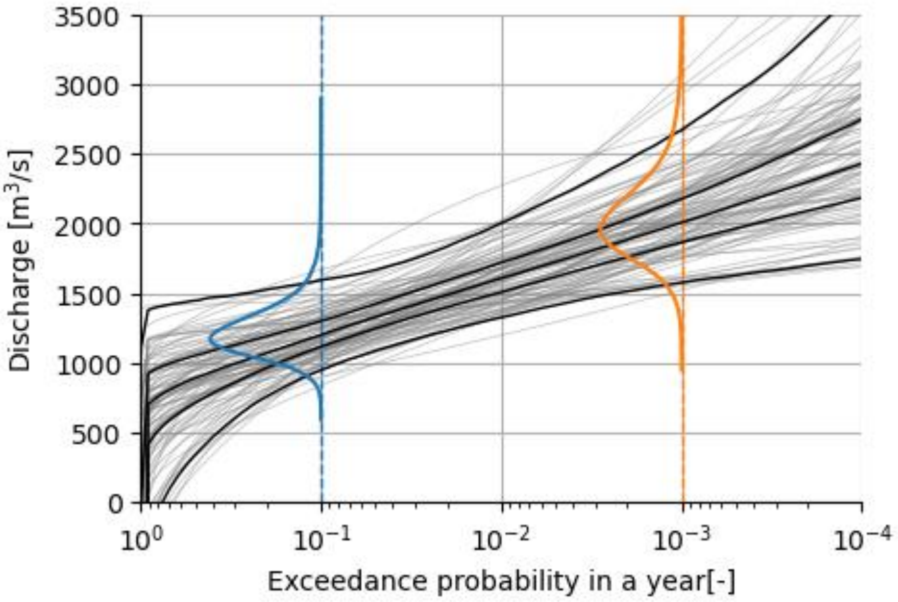
line 154: are the 26 questions made available somewhere?

These will be added to the supplementary information.

line 173: the weakly informed prior in Appendix A is very peculiar to me. I imagine very strange parameter combinations, very far from what could be expected for floods, are given the same weight than more reasonable ones, and some reasonable ones are excluded because of the bound at 10000. Why not the usual priors for the GEV distribution when dealing with floods, i.e., unbounded uniform for location and for the log of the scale and the Martins and Stedinger (2000, doi:10.1029/1999WR900330) geophysical prior, or similar ones, for the shape parameter?

The initial reason for choosing the weakly informed prior was that an uninformed prior did not result in uninformed 10-year and 1000-year estimates, which is what we were looking for. It gave bad results for the EJ-only results, as the 10-year and 1000-year distribution leave too much freedom for the uninformed GEV-parameter space. Aiming for uninformed 1000-year estimates evolved the approach to the one described in appendix A, in which a prior estimate for the shape parameter and a custom ‘empirical’ copula were used.

While the approach works and gives the required results, after reading this review, we consider exchanging the prior for the shape-parameter to the beta(6, 9) between [-0.5, 0.5] from Martins and Stedinger (2000). A quick analysis shows that it constrains the parameter space enough to give good results for the expert-only data, as shown in the figure below. This result is similar to Expert A, French Meuse in the supplementary material. Doing so will remove the need for appendix A, and make the approach less ad-hoc.



lines 185-195: here the expert information is accounted for as data (part of the likelihood). Why not accounting for it as prior information? That would be the natural way to do it: since the experts give their estimates without using discharge data, this can be considered prior information. For getting the prior distribution of the parameters from the prior assessment of the quantiles, one could use the procedure described in Renard et al. (2006, doi:10.1007/s00477-006-0047-4), for example. This would avoid the subjective choice of weights presented in lines 196-205, which actually control the fit of the tail of the flood frequency curves. Also, this would provide a more defendable prior than the one discussed in Appendix A.

As discussed in item 2 in the beginning of this document: considering the expert estimates as prior information and updating them with data, makes that the fitted distribution tend to follow the data more than the expert judgment. From a Bayesian perspective this is what you'd want, however it suppresses the effect of the expert's judgements in this EJ study. For that reason, we rather have the option of specifying the ratio between the two (as Eq. 6 provides).

This issue can also be viewed from the choice for the GEV distribution. The measured flows from the individual catchments could be considered independent identically distributed (iid). Not completely, due to climate change, and catchment changes, and human interactions (compromising the identically). But when combined with the expert estimates, this is further compromised. Therefore, instead of choosing a factor, we could have considered using a different or a mixed distribution that naturally fits the combination of data and estimates (but which?). We'll add this to the discussion.

line 196: log-likelihoods are summed

Indeed, will be corrected.

line 206: please indicate in which equation (and with what symbol) the "factor between the tributaries' sum and the downstream discharge" has been introduced. Is it the one in Eq. (1)? And what are the observations to which a log-normal distribution is fitted? I am confused here.

$f_{\Delta t,u}$, in Eq 1. We will add this. For the data-only model, an estimate for this factor was needed as well. For this, the historical factors were calculated and a log-normal distribution was used to parametrize this (it fitted well and is non-negative).

Section 3.4: I am sorry but I don't understand the procedure at all. I wish I could suggest how to improve points 1 and 2, but I can't figure out what they do mean.

We will clarify this. The main catch here is that (in our model) the dependence model does not only determine the strength with which each tributary's GEV-realization correlate during an event, but also how the GEV-distribution (within their tributary's bandwidths) are correlated.

If we would not have a bandwidth, but just a single GEV-parameter combination per tributary, it would be relatively easy: draw a [9 x 10000] sample of multivariate normal realizations and convert this to the GEV-space. Sum the discharges and multiply with 10000 realizations of the factor. This gives 10000 downstream discharges from which the exceedance probabilities of interest can be calculated. (Note that this approach is repeated 2000 times)

We do however have a bandwidth of tributary discharge statistics from the MCMC trace, and would like to process this into the final answer. The bandwidth in the tail is mainly the cause of not having a lot of realizations in the tail (i.e., not having thousands years of data) We can reason that if one tributary would have a GEV-combination that gives a curve on the upper end of the bandwidth, it would be likely because of a high discharge event that its neighbour has experienced as well. The neighbouring tributary would therefore likely also have an 'upper' realization. This effect is modelled by correlating the GEV-realizations with the dependence model as well. So we draw a [9 x 2000] sample (9 tributaries, 2000 curves were generated), and transform these from normal space, to uniform space (cdf), to uniform [1, 8000] space. This gives an index of the GEV-parameter combination to choose (the GEV-curves will cross, so they were ordered based on their 1000-year discharge).

Lines 269-278: here it seems evident to me that the objective assigned to the expert is to guess a reasonable mean annual peak discharge, in m^3/s, but not so much the shape of the growth curve. Afterwards, the Cook's method values the experts in how well they get the order of magnitude of flood discharges right, more than the shape of the distribution. Is this what we need to inform our analysis about how extreme can large floods be?

The lines 269-278 describe the different approaches that experts chose to structure their reasoning. We tried to steer the experts as little as possible in choosing their approach (while you could argue that providing a certain type of data already does). All tried to structure their thinking in some way, some found an approach as calculating a mean annual peak discharge (or 10-year flow in our case) apparently most suitable, but we did not assign them a method.

Finally, it comes down to trusting that an expert's performance in estimating a 10-year flow gives a good indication of the expert's ability to estimate a 1000-year flow (and building on this trust by processing the results and comparing it with the Borgharen benchmark). There was no distinction between the importance of the return periods (such as, calculate mean annual flood and scale this to an extreme, would be). You can argue that estimating a growth-factor focusses the experts' attention more on the changes in catchment hydrology (or how big an extreme flood can be), but we prefer to directly assess the variable of interest (the 1000-year discharge) as it gives less ambiguity in processing the answers.

Line 290: "not too steep"

Will be corrected.

Figure 5: if I have understood well, the points in the third column should all be grey because discharges at Borgharen are not used in the fit. Am I right?

They should indeed be grey, thanks for the suggestion.

Line 308: I don't get what the following sentence means: "Sampling from these wide uncertainty bounds will therefore (too) often result in a high discharge event".

Will clarify or remove (perhaps this information is not too critical). What it means: When only fitting to data, the marginals will have very large uncertainty in the tail. Some sampled tributary discharges will have extreme (e.g., >5000 m^3/s) discharges, even before combining with the other tributaries. With 9 sampled tributaries, there is a reasonable change that a combination has one of these samples, causing the end-result to be pushed up. Nothing too strange, just what we found when looking into the results.

Figure 6bc: it seems peculiar that combining the pieces of information that individually result in the blue and yellow distributions leads to the red one (e.g., the red mode is lower than the blue and yellow ones). Can you comment on that?

The difference is that the yellow results used EJ for both 10 and 1000-year, while the red only for 1000-year. The difference in mode order is due to the curves tilting a bit because of the 10-year EJ estimate (more clear in Figure 5ghi).

line 323: why are the median values considered best estimates?

50% chance it is lower, 50% chance it is higher. But this ambiguity is why it is between apostrophes, the most likely value could be seen as ‘best estimate’ as well. We’ll remove it, a word like median probably doesn’t need clarification for the readers of this article.

line 330: I don't understand the sentence.

“it can point out the statistically accurate experts to improve the extrapolated range”: This means that the data can be used (in Cooke’s method) to determine expert performance, assign them scores, and use these scores for to adjust the extrapolated range (in comparison to what it would have been without EJ). We’ll replace ‘improve’ with something more neutral like ‘adjust’.

line 340: but the experts knew about the 2021 event when doing the exercise and this has biased their estimates, I guess. How would have their estimates been different before 2021? That's hard to tell.

Probably lower as well, given that it is generally perceived as an unexpectedly large event. But then again, mainly for the summer season which was not considered separately in the estimates.

line 350: the following sentence doesn't mean anything to me: "were combined ... in ranges that are commonly 'in sample'".

Will remove “in ranges that are commonly 'in sample'".” part (it seems that it should not be there).

line 360: since the tails of the distributions are controlled by the expert opinions, it seems to me obvious that they "seem credible". Couldn't they be compared to the outcomes of a more classical regional flood frequency analysis?

The tails are compared to the discharge at Borgharen, which is calculated from their estimates (and can be validation reasonably well because of the long data record). So while this is obvious for the tributary estimates, the final discharges at Borgharen (as shown in the supplementary information) do not have to seem credible.