Dear Mr. Viviroli and reviewers,

Hereby we submit the fourth revision of the article: “Using structured expert judgment to estimate extremes: a case study of discharges in the Meuse River”. We thank the editor for reconsidering the article, and thank the referee 2 for reviewing the article again.

Compared to the last revision, the changes specifically address the comments of referee 2, which mainly concern the use of structured expert judgment in the hydrological context of this study. We now state specifically that we apply the Classical Model for experts judgment, but without evaluating the assumptions in this method, as this has been extensively done in (referenced) literature. We also mention that our method does not replace of supersede a more typical hydrological or hydraulic modelling approach, but that it can be used as an alternative to estimate uncertainties in extremes. We also adopted the referee’s suggestion for improving Figure 5.

A detailed response to the comments is found on the next pages.

Together with this document, we uploaded:

- The new version of the article.
- A comparison between the old and new version using track changes. Note that in addition to the items mentioned above, we revised the full article again and made some minor changes while performing our review. Line numbers or section references are added in the response to the comments to trace where the comments have been processed.
- The (unchanged) supplementary information.

We hope that the manuscript changes clarify the Classical Model for structured expert judgment to the referee and potential readers. We thank the referees and editor for their effort and input, and hope that the changes following from their comments have made this work into an appealing article for the hydrological community.

Kind regards,

Guus Rongen
Oswaldo Morales-Nápoles
Matthijs Kok

RESPONSE TO REFEREE 2'S COMMENTS (presented in report 1)		
#	**Referee comment**	**Authors' response**
1	Regarding the authors' reply in the 1st comment of the previous review: If the experts perfectly estimate one of the three (i.e., 5%, 50%, 95%) quantiles of the 10-year discharge for each tributary, In my opinion, the actual values of the percentiles should be shown/compared, and not what is shown in Fig. 4, which is confusing. First of all, for what percentile (5%, 50%, 95%) are the estimates are shown in Fig. 4 entitled "Seed question realizations compared to each expert's estimates". This Figure shows the uncertainty/distribution (of the 5%, 50%, or 95% percentile) as constructed from 70 values (7 experts times 10 estimates per tributary)? If yes, is it correct to construct the distribution (of the 5%, 50%, or 95% percentile) from all the estimates while some seem to be completely off (i.e., with the exception of the expert D and maybe E, the rest experts seem to have estimates with a low probability of occurrence based on the constructed distribution).	Figure 3 is meant to give a visual representation of the concept of "statistical accuracy" or "calibration" in Cooke's sense. In order to clarify this we have included further clarification related to the calibration score. In lines 178-179 we add "(the quantity $2 \cdot N \cdot \sum i=1,...,4 \; si \log(si/pi)$ is asymptotically χ_3^2)". In lines 182-184 we added *"Figure 4 is presented to visualize the disagreement between si and pi for this study. This figure will be further discussed in subsection 4.1. For now, it is sufficient to note that the agreement between si and pi is highest for expert D"*
2	Regarding "Because the goal is to elicit uncertainty, experts estimate percentiles rather than a single value. Typically, these are the 5th, 50th, and 95th percentile.", why not have asked them to also estimate the mean and variance, which are very useful (for example, why calculate the 10-year estimate from these 3 quantiles and through a distribution fitting rather than ask the experts to give at least the mean of their estimates)?	Cooke's method for structured expert judgment is based on the elicitation of quantiles from expert's uncertainty distribution. Other methods may be based on the elicitation of other quantities (moments for example). Investigating those alternative methods is out of the scope of our research. We changed the title of section 3.2 (line 157) to better reflect this to *"Assessing uncertainties with the Classical Model for expert judgments"*
3	3) For using the Metalog distribution, the authors state that "This distribution is capable of exactly fitting any three percentile estimate.", but many flexible 3-parameter distributions can be fitted by 3 estimates. Similarly for the ratio, where the log-normal distribution is fitted, I think that these distributions should be used in caution (for example in expressions like "as it is unlikely that the 1,000-year discharge is lower than the highest on record"), since they may confuse the readers thinking that these are the actual distributions estimated in this study for the percentiles and discharge ratios, whereas only a few data are used for the fitting and thus, they do not capture other attributes of the distributions (e.g., its tail, etc.; for example, it is shown that streamflows follow a heavy-tail distribution, and thus, the discharge ratio should have a similar tail definitely heavier from the log-normal's one).	We've added extra clarification as follows: *"Notice that for this research, the Metalog distribution represents the uncertainty distribution of each expert over a particular discharge with a given return period. While it is related to the underlying distribution of extreme discharge it does not make any assumption about this underlying distribution other than the ones expressed by experts through their percentile estimates"* In lines 209 – 212 Regarding the log-normal distribution for the ratio (downstream discharge divided by upstream discharge), we added extra clarification as well: *"The ratio itself does not represent streamflow, so there is no need to assume a heavy tailed distribution as would be expected for streamflow (Dimitriadis et al., 2021)"* In lines 291-292. The elements that contribute to these ratio are explained in Section 3.1, lines 133-137.

4	4) I am also concerned about the assumption “An implicit assumption is that the experts’ ability to estimate the seed variables (a 10-year discharge) reflects their ability to estimate the target variables (a 1000-year discharge).". The 10-year discharge is a not-so-extreme value, while the 1000-year discharge is considered extreme. It has been shown that streamflows follow a heavy-tail distribution (see, if found useful, the largest performed global analysis in Fig. 11 of https://www.mdpi.com/2306-5338/8/2/59, where streamflow is shown to be almost as heavy-tailed as precipitation, which is known to follow Pareto-tail as indicated and extensively discussed in https://www.itia.ntua.gr/en/docinfo/2000/), and so, an expert may have a rainfall-runoff model that is good only in estimating regular discharges rather than extreme ones (or the other way around) that require a separate rainfall-extreme analysis (since the 1000-year rainfall cannot be easily estimated from the observations). I would recommend reflecting on this issue in the Abstract, Conclusions, and maybe even the Title.	*We have modified the text to make it clear that the purpose of the paper is not to reflect on the underlying assumptions of the Classical method but rather to discuss it’s potential in improving hydrological studies of extremes. We changed “This assumption is in fact one of the most crucial assumptions in the Classical Model and has extensively been discussed in, for example, Cooke (1991).”* To *“This assumption is in fact one of the most crucial assumptions in the Classical Model. The objective of this research is not to investigate this assumption. For an example of a recent discussion on the effect of seed variables on the performance of the Classical Model the reader is referred to Eggstaff et al. (2014). The representativeness of the seed variables for calibration variables has extensively been discussed in, for example, Cooke (1991).”* In lines 496-499
5	5) Regarding the "However, an informative prior was added to the shape parameter because, with only expert estimates and no data, two discharge estimates are not sufficient for fitting the three parameters of the GEV-distribution. Additionally, the variance in the shape-parameter decreases with increasing number of years (or other block maxima) in a time series. The 30 to 70 annual maxima per tributary in this study are not sufficient to reach convergence.". These are all discussed and analyzed in Koutsoyiannis 2004 (a,b), where it is suggested (Fig. 5-6 in 2004a and Fig. 10-11 in 2004b) that small sizes of records, e.g. 20–50 years hide the distribution's EV2 shape parameter around 0.15 + 0.05 (e.g., in Fig. 13 of 2004b, as estimated from only the largest-lengthed precipitation records above 100 years). D. Koutsoyiannis, Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation, Hydrological Sciences Journal, 49 (4), 575–590, doi:10.1623/hysj.49.4.575.54430, 2004a. D. Koutsoyiannis, Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records, Hydrological Sciences Journal, 49 (4), 591–610, doi:10.1623/hysj.49.4.591.54424, 2004b.	We thank the reviewer for pointing out these references. They have been added to our paper. In lines 269-270 we write *“Similar observations have been presented before for extreme precipitation in Koutsoyiannis (2004a, b)”*
6	6) It is mentioned that "When estimates on uncertain extremes is needed, which cannot satisfactorily be derived (exclusively) from a (limited) data-record, the presented approach provides a means of supplementing this information. Structured expert judgment provides an approach of deriving defensible priors, while	We thank the reviewer for his/her kind words regarding respect for our work. Similarly we respect the reviewers work and have tried to reply to his/her comments the best way we can. We appreciate the reviewers observation that different modelers would try to approach the same problem differently. We agree that professionals will try to use the best tools

	the Bayesian framework offers flexibility for incorporating these into probabilistic results by adjusting the likelihood of input or output parameters.". However, when estimates on extremes are needed, one requires the best statistical approaches in the literature (if direct streamflow records are available) or some, equivalently robust, rainfall-runoff models (if only rainfall records are available) that can capture several hydrodynamic aspects of the selected area (as explained in my previous reviews). From either approaches, one can then estimate the uncertainty of the results from these approaches or models. This is not equivalent (and should not be confused) with some experts using (different or even the same) statistical approaches or models in a robust (or maybe incorrect) manner. Additionally, I would follow a more traditional approach, and see which of the expert(s) seem to achieve (in general or for each tributary) better performances in their predictions (which would mean that they have a better understanding of the area and their applied models/methods), and I would follow their suggestions and not the ones from the rest of the experts that they did not perform well. I respect the authors' work and I would appreciate their reply to this, which is at the core of their paper.	at their disposal be it models, data collected from the field, experiments (when available) or expert judgments. Our paper provides a well-executed instance of the Classical Model for expert judgments for estimating uncertainty regarding extreme discharges. We show that a well-executed instance of the Classical-Model combined with Bayesian inference can be one of what researchers may regards as the best tools at their disposal. We don not claim it is the only one and we have modified the conclusion to reflect this. We change “*When estimates on uncertain extremes is needed, which cannot satisfactorily be derived (exclusively) from a (limited) data-record, the presented approach provides a means of supplementing this information. Structured expert judgment provides an approach of deriving defensible priors, while the Bayesian framework offers flexibility for incorporating these into probabilistic results by adjusting the likelihood of input or output parameters*” to “*When estimates on uncertain extremes are needed, which cannot satisfactorily be derived (exclusively) from a (limited) data-record, the presented approach provides a means (not the only mean) of supplementing this information. Structured expert judgment provides an approach of deriving defensible priors, while the Bayesian framework offers flexibility for incorporating these into probabilistic results by adjusting the likelihood of input or output parameters.*” In lines 554 – 558. We changed “*In our application to the Meuse River, we successfully elicited credible extreme discharges. However, a case studies for different rivers should verify these findings. Considering the credible results and the relatively manageable effort required, the approach presents an attractive alternative for complex hydrological studies where the uncertainty in extremes needs to be constrained.*” To “*Our research does not discourages the use of more traditional approaches such as rainfall-runoff or other hydrodynamic or statistical models. Considering the credible results and the relatively manageable effort required, the approach (when well implemented) can present an attractive alternative to models that approach uncertainty in extremes in a less transparent way.*” In lines 558 – 562.
7	7) In Figure 5, please indicate the observed/fitted 50th percentile of the 10-year and the 1000-year (through fitting model) discharges to compare with the experts' estimates.	We have adopted this suggestion and added the 10-year and 1,000-year discharges as derived from the data (the 50th percentile) to figure 5. This illustrates the relative proficiency of expert D and E in estimating discharges (even though it is not about the median, but about the full uncertainty estimate).