Dear Mr. Viviroli and reviewers,

Hereby we submit the updated version of the article: “Using structured expert judgment to estimate extremes: a case study of discharges in the Meuse River”. We took the feedback of the reviewers and editor to heart and made significant changes to the article. The article now has a more general focus on estimating (hydrological) extremes rather than being specifically tailored to the Meuse’s extreme discharges. While the latter is still the article’s case-study, the following changes should make the research more appealing to a broader public of statistically focussed hydrologists:

- Cooke’s method and structured expert judgment have been given a more proper introduction: How it compares to and formalizes regular expert judgment.
- The Bayesian approach is more general now (thanks to reviewer 1’s feedback). The ad-hoc prior and the extra weight for EJ-likelihood are removed. This should make the approach more easily applicable (and therefore more relevant) for other studies.
- In line with this, the abstract, introduction, discussion, and conclusion now reflect on the study as a method to estimate ‘out of sample’ extremes in general, and less on the comparison between the study’s results (i.e., using data, using EJ, or using both).

The next page contains an overview of the main changes made to the manuscript, ordered by section. A detailed response to both referees’ comments is found on the pages thereafter. These responses have been updated from the response we had given during the open discussion phase.

Together with this document, we uploaded:

- The new version of the article.
- The new version of the supplementary information, now also including the questionnaire through which the uncertainties were elicited.
- A comparison between the old and new version using track changes. Note that large parts of the text have been changed, which makes it hard to track the differences. Therefore, we included line numbers in the response to the comments (the tables hereafter) to indicate specifically where the comments have been processed.

We think that the manuscript changes are a substantial improvement and hope that the new version will be reconsidered. However, regardless of the decision, we would like to thank the reviewers for their effort and input so far, as their feedback has greatly improved the presentation of our research.

Kind regards,

Guus Rongen
Oswaldo Morales-Nápoles
Matthijs Kok

Overview of the main changes per section

1. Introduction
   - GRADE is now clearly presented as the benchmark in this study and named as a regional flood frequency analysis (in response to referee 1's comments).
   - We are no longer comparing the presented 'data-based' approach to a 'model-based' approach (in response to referee 2's comments). This presents the choice as binary, while lots of hydrological models are a combination. We now present several approaches of extending a data record, and present expert judgment as an alternative to these.
   - Structured expert judgment is introduced more properly and is compared to regular 'everyday' expert judgment (in response to referee 2's comment). This should make the exact meaning of it more clear to a broader audience.
2. Study area and used data.
   - A list of used data is given (in response to referee 1's comments)
3. Method description
   - Section 3.2: Cooke's method is introduced more clearly. What it is, why you should use it, and specifically the *structured* part.
   - Section 3.3: This section is largely rewritten, to match it with the changes in the Bayesian approach (i.e., the geophysical prior and removing the EJ-factor). The explanation is more formal as well (i.e., includes a proper mathematical description).
   - Section 3.4 is simplified (in response to referee 1's comments) and is accompanied by a new appendix (A) that gives a mathematical description of the algorithmic steps used to sample downstream discharges.
4. Results
   - The result sections did not change significantly apart from some extra explanation on the downstream discharge results in Section 4.4.
5. Discussion
   - This section has been rewritten completely. It is split into two parts, method-related (5.1) and result-related (5.2). Section 5.1 explains why we elicit discharges instead of ratios or shape parameters (in response to referee 1's comments), the choice and suitability of the GEV distribution, and the omission of seasonality. Section 5.2 discusses the validity of results, it
     - reflects on bad GEV fits and bad Cooke's Method scores (which correlate),
     - reflects on the comparison with GRADE,
     - compares estimation of extremes through EJ (with Cooke's method) to extrapolation with a model, and
     - explains the EJ-only approach (without using data) as too uncertain.
6. Conclusions
   - The conclusions now focus less on comparing the combination of data and EJ to data-only and EJ-only, as that is less relevant for general applications.
   - The conclusions contain a general statement on how expert judgment can be used to limit uncertainty through a Bayesian approach, because this is an important 'take-away' message for the readers.

RESPONSE TO REFEREE 1'S COMMENTS		
#	**Referee comment**	**Authors' response**
1	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.	The proposed method is indeed tailored to get the flood magnitude right. However, we do so by applying Cooke's method for Structured expert judgment, in which the experts are scored based on their ability to estimate the 10-year discharges. These are then used to estimate the 1000-year discharge. Consequently, experts that score high will have a 10-year estimate that corresponds to the observations, and a defendable estimate for the 1,000-year. Together, these indicate the ratio, or shape (if combined with data) between less extreme and more extreme discharges. We chose to elicit discharges rather than a ratio or shape parameter, as it directly informs our quantity of interest. Indirectly, these parameters are thus derived from the 10-year and 1000-year estimates. Eliciting such parameters directly could indeed change the focus more to a comparison between events of different extremity, which is now discussed in Sect. 5.1, lines 445 – 459.
2	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.	We thank the reviewer for this suggestion, which has greatly improved the presentation of this research. We have changed the approach for Bayesian inference in three ways (see article Sect. 3.3): - A geophysical prior is used for the shape parameter. This replaces the ad-hoc prior from the old Appendix A, making the approach simpler and more defendable. - The expert estimates are considered priors now (additional to the geophysical shape-parameter prior). This is mainly a matter of wording, the contribution to the posterior likelihood is unchanged, except that: - The weighing factor between expert judgment and observations is removed (including Appendix B, which showed its sensitivity). Rather than trying to fix deviating expert estimates, we now use it as an extra check if the experts' estimates are plausible (Experts D and E, with high scores, have 1000-year estimates that align with the observations, through the GEV). Regarding this specific comment: we do now consider the expert estimates a-priori information (see lines 215, 532).

3	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.	As described under the response to the last comment (2), we have removed the factor. Estimates that do not match the observations through the GEV-estimates can in some cases be defendable. For example, when an extreme would no longer be considered to be from the same population (i.e., identically distributed) compared to less extreme events. However, the GEV is generally flexible enough to facilitate light and heavy tails. So, rather than fixing deviating expert estimates, we accept them and use them as a check of whether the high-scoring experts have estimates matching the GEV. This is the case, which also means that 'bad' fits are filtered out through Cooke's method. See the discussion on using the downstream discharges as check in lines 438-444, and the further discussion on the GEV as validation in line 460-469, and 485-489. Throughout the text, we have clarified that we estimate upstream discharges, and calculate downstream discharges.
4	4) the results are not assessed against a benchmark. Why not using regional flood frequency analysis as a benchmark?	We have added some more explanation on GRADE. This model (Generator of Rainfall And Discharge Extremes) a variant on a conventional regional flood frequency analysis that includes historic events. It resamples historical rainfall and simulates this with hydrological models (instead of estimating discharges from statistical catchment properties). We clarify in the latest version of the paper that our results are assessed against this benchmark: Line 36-39 introduce GRADE as regional flood frequency analysis. Line 82 explains the comparison to GRADE (benchmark), which is further presented in line 414 onwards. GRADE's suitability as comparison is discussed in line 490-499
	5) some of the methodological steps are unclear, sometimes, and should be properly explained (see the detailed comments below).	See below
5	line 8: MCMC is just a tool. I would say here that you use Bayesian inference.	Changed (line 8)
6	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 m3/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. This methodological choice is now discussed in the discussion (Sect. 5.1, lines 469-474).

7	line 30: here the text suggests that hydrological model simulations outperform statistical methods in flood frequency analysis. Has this been demonstrated in the literature? **@1** 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? **@2**	**@1**: Not as far as we know. We have removed the part in the introduction that suggests it (which we did not mean to do) and made it less "models vs. statistics". **@2**: Like mentioned in the answer to comment 4, we now present GRADE as (a variant to) a regional flood frequency analysis, which it is. We have clarified that the results are compared to this (next to observed discharges), and Fig 7 (previously Fig. 6) now mentions GRADE as well, instead of "WBI-statistics", which is the same but was not introduced.
8	line 43: I don't get the factor 3 vs. 1.4 sentence. What is the "outcome"?	We have removed this sentence and the reference.
9	lines 65-68: spoiler alert! I would move this sentence after the results section.	We have removed this paragraph
10	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 experts' 10-year estimates to data). We have moved the sentence down to the method section that explains the sampling method (line 301-302). Here it becomes relevant (why we sample from 9 instead of 10 tributaries). The reader should at this point have enough background to better understand it.
11	line 85: I would add a table here in the main text summarizing the data provided to the experts.	We have added a list with a short description of the provided data in Sect. 2 (lines 103-115).
12	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.	We agree with the reviewer's observation. However, publishing the dependence results before these would render similar problems. 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).
13	line 109: Each variable is modelled by a marginal distribution, it is not a distribution.	This line is no longer present in the revised article.

14	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.	Please refer to comment 1 on why we chose to elicit 10-year and 1000-year discharges. Regarding the discussion on assessing the weight of an expert based on their ability to estimate a ratio rather than an absolute value: that is a valid point. It would indeed be interesting to compare the results with a study focused on that. We have added this to the discussion (lines 445 – 459).
15	line 151: "a training exercise"	Corrected (line 201)
16	line 154: are the 26 questions made available somewhere?	They have been added to the supplementary material
17	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?	As mentioned in the response to comment 2, we have changed the old weakly informed prior by the geophysical prior from Martins and Stedinger (2000, doi:10.1029/1999WR900330). See lines 239 – 248. It is indeed a much more straightforward way of limiting the shape-variability of the GEV. For the location parameter, we used a weakly informed prior that gives a uniform likelihood for all positive values (-inf for negative flows which can safely be assumed infeasible in our area of application). For the scale parameter we used a uniform distribution for positive values as well. Note that we did not use the Jeffrey's prior (i.e., 1/scale uniform), as this gave bad results in combination with the expert estimates and without data: The very high probability density of scale values close to 0 would result in a more or less horizontal GEV curve: a discharge for the location parameter that seems plausible given the 10-y and 1000-y expert estimates, and a near-zero scale. The prior is discussed in lines 237-252.

18	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 mentioned in the response to comment 2, we now consider expert estimates to be prior information. We incorporate this into the posterior log-likelihood function with the method presented in Viglione et al. 2013: DOI:10.1029/2011WR010782. We find this a straightforward and easy to implement procedure for incorporating expert judgment. Note that the subjective choice of weights, previously discussed in 196-205, was not needed because of this procedure. It was due to the likelihood of the observations being dominant compared to the prior likelihood. We suspect (but haven't checked this), that this would be similar when following a different procedure such as mentioned by the reviewer.
19	line 196: log-likelihoods are summed	Corrected (line 266).
20	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.	$f_{\Delta t}$, in Eq. 1. We have added this (line 274-275). 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). This text in lines 274-286 was adjusted to make this procedure clearer).
21	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 have clarified this procedure. Section 3.4 now contains a conceptual explanation, and (the new) Appendix A has been added to give a step-by-step overview of the calculation. Together, we trust the method will be clearer to the reader.
22	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?	Please refer to the responses on comment 1 and 14. We appreciate your suggestion of estimating ratios and assessing experts by their ability of estimating ratios, which we've added to the discussion.
23	Line 290: "not too steep"	Corrected (line 383)
24	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?	We have made these grey, as well as the observation dots in the supplementary information for the Borgharen discharges. (See Figure 6 (prev. 5) and the supplementary material)

25	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".	To give more background on this: when usually fitting a model, the whole model is fitted to the end result. For the sake of the expert elicitation, we fitted it to individual components, just like the expert were estimating. When sampling from these components, the model-simplifications (e.g., the Gaussian copula does not exactly reproduce tail dependence, or the fitted GEV does not exactly reproduce the tributary discharges) result in a slight deviation.
26	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?	This is because the red line results from the GL DM, and the yellow line from the EQ DM. The latter has a higher 'factor between upstream sum and downstream'. This is now explicitly explained in lines 402-405.
27	line 323: why are the median values considered best estimates?	The term 'best estimates' is removed. Median should suffice for the readers. (Line 417)
28	line 330: I don't understand the sentence.	The sentence is changed for: "Including expert estimates, weighted by their ability to estimate the 10-year discharges, improves the precision of discharge estimates in the range of extremes." Note that while Cooke's method gives a defendable way of saying that the *accuracy* improves, we stick here to the safer statement of saying the *precision* (i.e., the narrowness of the uncertainty intervals) improves (line 433).
29	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.	Indeed, it most certainly did affect their estimates, and it would most likely have affected the comparison to GRADE as well (if GRADE would have included the 2021 event). This is now more elaborately discussed in the comparison to GRADE in the discussion (lines 490-499 and specifically line 497).
30	line 350: the following sentence doesn't mean anything to me: "were combined ... in ranges that are commonly 'in sample'".	Changed for: "Experts' estimates of tributary discharges during a once per 10 year and once per 1,000-year event are combined with high river discharges measured over the past 30-70 years." (Lines 518 – 520)
31	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?	Please refer to the response to comment 4, GRADE can be considered a proper regional flood frequency analysis. Moreover, the expert did not estimate the downstream discharges directly, so their knowledge of these discharges could not directly inform their estimates (regarding tributary discharges).

RESPONSE TO REFEREE 2'S COMMENTS		
#	**Referee comment**	**Authors' response**
1	1) The main point raised by the authors for someone to use the suggested method is that "...existing statistical and hydrological models that estimate these discharges often lack transparency regarding the uncertainty of their predictions..."; however, please note that the purpose of the probabilistic analysis is exactly this one (i.e., to estimate and take into consideration the uncertainty and variability of predictions of the input and output parameters of a flood model; see for example, a review, applications, and discussion on the uncertainty of flood parameters through benchmark examples in Dimitriadis et al., 2016). I would suggest not comparing with such methods (which are plenty in the literature), but focusing on the advantages and limitations of the proposed method.	We have rephrased parts of the introduction to remove to the model-based versus statistics-based suggestion. Like suggested by the reviewer, we focus on the advantages and limitations of using structured expert judgment, to reduce the uncertainty in extreme discharges. Some other approaches of doing this (e.g., paleoflood, historical archives) are presented as a comparison in the introduction now (lines 51-55), rather than a comparison to model-based approaches. The (mostly new) discussion Sect. 5.1 discusses the advantages and limitations of the proposed method.
2	2) The fact that "...the devastating flood event that occurred in July 2021... was not captured by the existing model for estimating design discharges.", is not for the statistical methods to blame (or replace), but a more appropriate analysis by experts should have been performed. For example, there is an application shown in Figure 10 (in Dimitriadis et al., 2016), where there was a certain flooded area that could not be captured by a 1D model (due to the 1D nature of the model that cannot account for a 180 degrees turn of the water, since only 1 direction is possible within a cross-section), whereas this area can be captured if a 2D (or quasi-2D) model is applied. However, only an expert in flood modeling could identify this (e.g., the authors state that "The study demonstrates that utilizing hydrological experts in this manner can provide plausible results with a relatively limited effort, even in situations where measurements are scarce or unavailable."). If this is what the authors are trying to highlight in this work (i.e., that the flood models should not be blindly applied by non-experts), then this is a strong and important statement, which however needs to be further discussed.	We agree with the reviewer that up to a certain point, the impacts of extreme events can be estimated better when a good analysis of the hydraulic details are made. For the July 2021 flood, a clear example of this is the effect of dams in the catchment (hydropower as well as weirs). The main cause of the event being surprisingly large was however the meteorological situation. Therefore, we do not think an appropriate analysis of the hydraulics would have led to the model (GRADE) capturing the event. The most important point here is that practitioners need to be aware of the uncertainties in their modelling approach, a point on which we think the reviewer and the authors agree, if we understand your comment correctly. Accordingly, we do now clearly present 'expert judgment' as the ability to use one's experience to verify observations (referring to the reviewer's "a more appropriate analysis by experts should have been performed"). And structured expert judgment with Cooke's method as way of formalizing expert judgment. Lines 56-60 introduce expert judgment in this context, and the discussion (both 5.1 and 5.2) now more extensively contains a description of how the participants performed in this regard.

3	3) Please consider rephrasing the sentence "Quantifying events that are more extreme than ever measured (i.e., with return levels that are longer than the time period of representative measurements), requires extrapolating from available data or knowledge.", since it is not exactly true. The return period T corresponds to a probability of occurrence (i.e., on average, a storm event is expected to occur in T years) and not a deterministic occurrence that involves any kind of extrapolations or specific (i.e., 5th, 95th etc.) quantiles (please see the mathematical definitions and methods for extreme analysis and probability fitting in a recent work by Koutsoyiannis, 2022).	The previous wording indeed suggest that historic data carry a deterministic return period that can be extrapolated like a data point. We have changed this for: " Estimating the magnitude of events greater than the largest from historical (representative) records is a nontrivial task. It requires establishing a model that describes the occurrence of such events and subsequently extrapolating to specific exceedance probabilities from this model." (Line 30-31)
4	4) The application of Cooke's method to the specific study is not very clear to me. For example, the authors state that "A simple statistical model was developed for the river basin, consisting of correlated GEV-distributions for discharges in upstream sub-catchments. The model was fitted to expert judgments, measurements, and the combination of both, using Markov chain Monte Carlo. Results from the model fitted only to measurements were accurate for more frequent events, but less certain for extreme events."; since they were all experts and applied the same model, how come they came up with different results, did they use different methods, and what are these methods? where did the experts base their reply, did they perform also simulations or just probabilistic fitting?	The participants in the study are all called experts in the article. They were pre-selected on their field of expertise (practitioners or researchers in hydrology). However, their expertise as uncertainty assessors is subsequently assessed using Cooke's method. Therefore, whether they are expert (in assessing uncertainty) in the context of this study is determined based on their estimates for the 10-year discharges (explained in Sect. 3.2, and explicitly in line 207-208). We have clarified this by giving Cooke's method a more proper introduction (line 56-73, and Sect. 3.2). Note that the model (Eq. 1), is a framework used by us to process the experts' estimates. The experts had to come up with 10-year and 1000-year discharge estimates (5th, 50th and 95th percentiles, such that it is an uncertainty assessment). The experts were free in choosing their methods to come up with the estimates needed to fill in Eq. 1. This was indeed unclear, so to clarify this, the first sentence of Sect. 3 is now (note the bold words that were added): *To obtain estimates for downstream discharge extremes, experts needed to quantify* ***different components in*** *a simple model that gives the downstream discharge as the sum of the tributary discharges, times a factor correcting for covered area and hydrodynamics.* The expert session was a 1-day expert session (lines 200-203), in which the experts had to come up with uncertainty estimates for 10 tributaries, which tends to steer them towards using simpler 'models' for making their estimates. The experts didn't have to do simulations or probability fitting but could so if they deemed it necessary. We have added this in Sect. 4.2. as well (lines 368-370).

5	5) In my opinion, it is not very appropriate to apply a Monte-Carlo method with so few samples; please consider including more samples. Also, how come "The combined approach provided the most plausible results, with Cooke's method reducing the uncertainty by appointing most weight to two of the seven experts."; why the authors have selected these 2 scientists; were these two more experts than the other scientists?	We have clarified Section 3.4. It is now split in a more conceptual part (Sect. 3.4) and a more mathematical part (Appendix A). We used 10,000 samples for each tributary, which are used to generate an exceedance frequency curve for the downstream location. These 10,000 yearly discharges are sufficient to cover up to the 1,000-year range. By doing this 10,000 times, we also get uncertainty bounds for this (10,000 was deemed sufficient for estimating the 2.5th, 35th, 50th, 75th, and 97.5th percentiles). We now clearly mention that the whole simulation comprises 100,000,000 samples (lines 306, 576), but split in 10,000 times 10,000 to create uncertainty bounds. Regarding the second part of the reviewer's question: The 2 experts were assigned the greatest weights based on their statistically accurate estimates for the 10-year discharges on the 10 considered tributaries. Within the context of this study (lines 207-208), we consider their uncertainty estimates more valuable. Saying that they are more experts than the others would have a connotation we wish to avoid. To clarify this, the sentence in the abstract (lines 10-11) was changed to: "Cooke's method reduced the uncertainty by appointing most weight to the two experts that could most accurately estimate more frequent discharges."
6	6) More details are required to back up the statement "The discharge at the Dutch border exceeded the flood events of 1926, 1993, and 1995. Contrary to those events, this flood occurred during summer, a season that is (or was) often considered irrelevant for extreme discharges on the Meuse."; please perform a proper statistical analysis and identify for each season the appropriate probability distribution to show at what discharge the probability of occurrence in the summer season exceeds the selected return period.	We added the exceedance frequencies presented in the (Force Fact-finding hoogwater, 2021) report to the article (already in the reference list): The corresponding author of this article did the EV-analysis for the discharges in that report, which showed that the flow had a 120-year average recurrence interval based on year-round statistics, and 600-year average recurrence interval when considering only the summer half year (April to September). Please refer to lines 22-25. These estimates are based on MCMC-fitted GEV-distributions including the 2021 event. Please refer to figure 2.5 in that Dutch report. We acknowledge that this is a Dutch report. An international publication closely related to this report is on its way but has not been published yet. If the reviewer would find it necessary, we see if we could include the EV-analysis in that report (which was done in a similar manner as the 'data-only' approach in this study) as an appendix in this article. We'd have to discuss this with the authors of the just mentioned yet to be published article to avoid duplication.

7	7) Regarding the comments "The event was thus surprising in multiple ways. This might happen when we experience a new extreme, but given that Dutch flood risk has safety standards up to once per 100,000 years (Ministry of Infrastructure and Environment, 2016) one would have hoped this to be less of a surprise." and "While most studies aimed at obtaining better estimates of discharge extremes use hydrological or statistical modeling, some follow the approach of using expert judgment (EJ).", please note that this is a must point in every scientific application, since when non-experts apply methods they do not understand, it could lead to failure regardless the magnitude of the selected return period.	As mentioned in the response to comment 2, we now mention that all modelling involves (or at least should involve) expert judgment to some extent. Subsequently we discuss how structured expert judgment with Cooke's method quantifies this process. See lines 55-60.
8	8) It is mentioned that "For the Dutch rivers Meuse and Rhine, the GRADE instrument is used for this. It generates 50,000 years of rainfall and discharges."; please give more details on this model and how it generates so long rainfall and discharge timeseries (does it use a stochastic simulation approach for the rainfall annual extremes and input these to a hydraulic model to produce the discharge at a specific location in the area of interest?).	The GRADE model is not scientifically published, but it is well described in this report: https://publications.deltares.nl/1209424_004_0018.pdf (Referred to as Hegnauer et al., 2014 in the article). We have added some more details on the GRADE method to the introduction of the article. Please refer to lines 35-44. Note that GRADE is the standard tool in the Netherlands (line 36-37), which is why we use it in this application.