Dear Mr. Viviroli and reviewers,

Hereby we submit the second 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 most of all thank the referees for reviewing the article again.

Compared to the first revision, the changes in this revision are concentrated in the presentation of Cooke's method (referee 2), Bayesian inference (referee 1), and the discussion of the study set-up (referee 2). The major changes thereby are:

- Cooke's method is explicitly presented as a method of estimating uncertainty, including how the experts are evaluated on their performance in estimating uncertainty (rather than the 'true' value. This clarifies Referee 2's comments regarding new/different observations, accidental good estimates, and more or different experts joining the project. Moreover, we refer to literature that researched the added benefits of using performance-based weights over equal weighting.
- The Bayesian approach is described with proper mathematical terminology, and the presentation (Sect. 3.3) is restructured to make the relationship with prior, likelihood, etc. clearer. Furthermore 5.1 discusses the (Renard et al., 2006) article.

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.

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 further clarify the study to the referees and potential readers alike. Again, we hope that the new version will be reconsidered, and thank the referees again for their effort and input, 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
   - Cooke's method (aka the Classical Model) is changed to Classical Model (aka Cooke's method). While they are two different names for the same method, the Classical Model is more consistent with recent scientific literature.
   - Structured expert judgment is introduced without referring to 'everyday' expert judgment.
   - Renard et al., 2006 is added to the examples of a study that uses EJ to limit inform extremes through prior information.
2. Study area and used data.
   - No changes
3. Method description
   - Section 3.1: The three components of the downstream discharge model (Eq. 1)_are listed explicitly, to distinguish the different models used in the study.
   - Section 3.2:
     - the Classical Model is introduced more clearly as a method for estimating uncertainty. The statistical accuracy is explained as (p-value based) method that scores the expert's ability to estimate uncertainty.
     - Literature that compares equal weighting to performance-based weighting out-of-sample is referenced.
   - Section 3.3: This majority of this section was rewritten of restructured to put it in context of Bayes theorem. The method is now addressed with proper Bayesian terminology.
4. Results
   - A comment on how experts are evaluated based on their ability to estimate uncertainty rather than proximity to observed values, in the context of Figure 4.
5. Discussion
   - Section 5.1: A more detailed comparison to Renard et al. 2006 is made, describing why one would choose that method (mainly: eliciting differences) over the method we chose.
   - Section 5.2: A notion is made of GRADE being method that is not published in a peer-reviewed scientific journal.
6. Conclusions
   - Minor edits

RESPONSE TO REFEREE 2'S COMMENTS (presented in report 1)		
#	**Referee comment**	**Authors' response**
1A	1) I understand the Authors' reply, but I still cannot comprehend what exactly is defined as "an expert's judgment". The traditional approach of an expert's judgment can be based only on models/methods. For example, the Authors mention that "Expert judgment (EJ), in terms of making estimates or verifying observations based on prior knowledge, is often unknowingly applied in everyday practice by researchers and practitioners"; however, in order to make an estimate one requires a model/method and historical observations for fitting/calibration, verification, and validation. Also, what do the Authors mean by "unknowingly"? An expert should know exactly what is (s)he doing and what are the impacts of the applied assumptions. …	Specifically regarding the sentence: *Expert judgment (EJ), in terms of making estimates or verifying observations based on prior knowledge, is often unknowingly applied in everyday practice by researchers and practitioners,* and *Also, what do the Authors mean by "unknowingly"*: We have removed this sentence (in the paragraph starting at line 57). It didn't refer to an "expert judgment" as commonly known, but to an everyday estimate or predictions that someone needs to make, which is confusing in this context.

1B	…Moreover, I do not entirely agree with this statement and Authors' approach for a different reason; even if the same expert applied different (equally justified) models to the same area (i.e., same case-study, initial/boundary conditions, same input, etc.), then it is certainly expected that the output would be different but not wrong (this is illustrated in the Dimitriadis et al., 2016 study). I think that this procedure is equivalent to the one where multiple (equally qualified) experts used different models/methods in the case study (which is my understanding that this is what the Authors illustrate in their study). However, even if a model/expert is closer to observation does not necessarily mean that this model/expert is better and should be assigned a larger weight coefficient, but (by assuming that all models/experts are equally justified/qualified) that there is an intrinsic uncertainty enclosed in different models/experts, which we should take into account in our flood risk management strategy rather than trying to narrow it down. The danger here is that in a future event, and since all models/experts are equally justified/qualified, the model/expert that was worst in the previous case-study could be now closer to the true observation, and therefore, would be wrong to have assigned a smaller weight coefficient. This would happen because after a limit the uncertainty is intrinsic and can be no longer removed/narrowed but rather only quantified, modelled, and considered in the management strategy. Please note that this is different than applying a wrong model/assumption in a case-study (as explained in my example in the previous review). …	Regarding the rest of the reply, and specifically: *then it is certainly expected that the output would be different but not wrong", and "However, even if a model/expert is closer to observation does not necessarily mean that this model/expert is better and should be assigned a larger weight coefficient, but (by assuming that all models/experts are equally justified/qualified) that there is an intrinsic uncertainty enclosed in different models/experts, which we should take into account in our flood risk management strategy rather than trying to narrow it down.*: First of all, we agree with the referee's comment: …*However, even if a model/expert is closer to observation does not necessarily mean that this model/expert is better and should be assigned a larger weight coefficient, …* We would like to underline that the Classical Model / Cooke's method does not evaluate experts based on their closeness to an observed value, but based on their ability to estimate uncertainty. Fig. 4 in the article illustrates this: expert D and E receive the highest weights because the quantiles/percentiles of their estimates represent the expected fraction (i.e., they are more uniformly distributed). In other words, if we consider an observed value to be randomly drawn from a distribution, and the expert perfectly estimates these distributions, the quantiles shown in Fig 4. will be uniformly distributed (or at least drawn from a uniform distribution). So Cooke's Method evaluates experts based on their ability to estimate uncertainty, and by doing so makes the method relatively insensitive to coincident in the observations. This is addressed with the changes: • Throughout Section 3.2 • Lines 350 and 351 Regarding the second part: *but (by assuming that all models/experts are equally justified/qualified) that there is an intrinsic uncertainty enclosed in different models/experts, which we should take into account in our flood risk management strategy rather than trying to narrow it down.* A recent study by Cooke et al., 2021 show the added benefit of performance-based weighting in an out-of-sample cross validation (earlier research was based on in-sample cross validation). Lines 202-204 are added to share the main finding of this study (increased informativeness without compromising statistical accuracy), which is the main reason for us to use performance-based weighting over equal weighting. With respect to the present article, we processed this comment by explained that Cooke's method scores experts based on the product statistical accuracy (SA) and informativeness, where SA is the dominant factor. This is because SA change significantly (orders of magnitude) across experts while informativeness is "more stable". A statistically accurate expert does not necessarily make estimates close to the seed question's answer (in fact often they don't), but makes estimates that represent the uncertainty in the answers. This is now clarified in two places in Section 3.2. First, with in the explanation of

		statistical accuracy, which is explained as being based on a p-value of statistical accuracy, and second: "*The statistical accuracy expresses the ability of an expert to estimate uncertainty. Because a variable of interest is uncertain, its realization is considered to be a value sampled from the uncertainty distribution. According to the expert, this realization corresponds to a quantile on the expert-estimated distribution. If an expert manages to reproduce the ratio of realizations within the interquantile intervals (such as in the example with 20 questions above), the probability of the expert being statistically accurate is high, hence they will receive a high p-value. Of course, this match could* be coincidental, like any *significant p-value from a statistical test. However, in general, a different sample of realizations (in this study, different observed 10-year discharges) is expected to give a p-value (i.e., statistical accuracy) of a similar order."* (lines 182 to 189) If an expert manages to capture the uncertainty well in their estimates, the statistical accuracy should be similarly high in case the realizations turned out to be different. Just like a statistical test for (e.g.) normality would give a p-value > 0.05 with 95% confidence, if another sample was drawn from a normal distribution.
1C	… If the Authors are certain that all scientists are experts, then I would recommend just quantifying the variability of their judgment/output (i.e., treating them as different 'models', equally correct and justified, as performed in the study by Dimitriadis et al., 2016), and assign an equal weight-coefficient.	In principle, we let the Classical Model decide what the expertise (i.e., uncertainty-estimating ability) of the participants (which we do call experts throughout the study) is, with regard to the questions in this study. However, it is common practice in the Classical Model to present equal weights (EQ) as well. In the main article the EQ results are only shown for the option in which no observations are used in fitting the results. The supplementary material however shows the full results for the EQ decision maker combined with observation (compare Fig 4.2 to Fig 4.3).

2	2) Regarding the reply to the 2nd comment, please present in a clear way what method/model/observations etc., has each expert used to derive his/her results, since "the ability to use one's experience to verify observations." is not a clear definition of an "expert judgment"; for example, what do you mean by "ability"? The only reason I can think of that one expert came up with a different output is that (s)he used different input, initial/boundary conditions, and/or methods in their thinking procedure (as explained in the previous comment). Specifically for the extreme analysis, if, by applying a method/model, the results constantly deviate from observations, then this would mean that the method/model is wrong, should be re-examined, and should be not taken into account in the management risk assessment through the Cookes method (in the recent book by Houstonians, 2022, there are plenty examples how one could severely underestimate the extremes if the assumptions are not correct, as in ignoring dependence, in assigning less robust or even invalid statistical estimators, in applying less accurate statistical distributions, etc.).	For clarity, there are two "models" to be distinguished: 1. The probabilistic model described in Section 3.1 (Eq. 1) that is used to calculate statistics of downstream discharges using Bayesian inference. 2. The models that each expert uses to calculate or estimate the components in Eq. 1. Model 1 is applied by us (the researchers), based on the experts' estimates from model 2. The performance of model 1 is evaluated in Section 4.4. More importantly, what approach the experts used for their estimates (the model 2) was asked during the expert elicitation and is described in *Section 4.2 Rationale for estimating tributary discharges*. We do not know exactly what models the experts have used. However Cooke's Method evaluates the methods used by individual experts (even when we don't know it exactly) based on the statistical accuracy of their answers (which we presume comes from their models). An expert which turned out to give very good estimates as evaluated by the combined score will be assigned a high weight regardless of the methods used (or not) in his/her quantification process. Similarly, if an expert applied a complex hydrological model which results in very bad estimates (for example constantly over or under estimating the seed variables), this indicates that the model is wrong, which results in the expert (and model) not being taken into account by Cooke's Method (as in the reviewer's example). To specifically address: *The only reason I can think of that one expert came up with a different output is that (s)he used different input, initial/boundary conditions, and/or methods in their thinking procedure (as explained in the previous comment).*: Differences in estimates would most likely results from differences in their rationale, as the experts were provided with the same information (presented in the supplementary information). Additionally, we know from their description of the applied method (presented in Section 4.3) that their approaches to answer the questions in the study differ.
4	4) But what are these components the Authors refer to in their reply and in the manuscript? This is important so that the Readers are able to criticize the experts' methods/models.	The components are now specifically listed at the end of Section 3.1 (lines 149 to 152). As explained in the previous item's response, we do not know exactly what methods were used to make estimates for these components, on top of what is presented in Section 4.3. While we agree that more detail on this would be an interesting addition for the readers, the study focus is not a hydrologic modelling study but on the ability of expert judgments to quantify uncertainty in hydrological problems. In the latter, the performance of their uncertainty estimates is what 'validates' their model.

5	5) But what if more experts join this project? More importantly, what if an expert's good judgment (i.e., closer to the true observation) was achieved by accident, and his/her assumptions no longer work for a future event where the conditions have changed?	*But what if more experts join this project?*: We do not know the effect of adding more experts. We do have two experts with a > 0.05 significance level, and with expert D having a SA of 0.683 it is unlikely that additional experts will strongly change the pooled result (but of course we can never know). For context, typically around 5 experts is deemed sufficient (Stephen, 2004 https://doi.org/10.1287/mnsc.1040.0205), just like 10 calibration questions (Colson and Cooke, 2017 https://doi.org/10.1287/mnsc.1040.0205). The robustness of the results is typically evaluated in a robustness analysis, in which sensitivity of the statistical accuracy and information score are calculated by leaving one or more experts out at a time and recalculating the measures of interest. The next two tables show the results from this, first for excluding experts, then for excluding items. This shows that the Global Weights DM is overall unsensitive to excluding one expert or item. Note that: • With respect to experts, the information score and SA are insensitive to specific experts. Note that the SA *increases* when removing the expert with the highest SA (expert D). Expert E (the second expert) becomes dominant, and the small influence of Expert G (third) corrects some of the over- or underestimates, leading to a higher SA. • With respect to items, the SA is relatively uncertain to specific items, except for excluding the Tabreux 10-year estimate. This is one of the items where Expert D scored significantly better than the other experts. Removing it shifts the weight more to Expert E, which for this particular case results in a lower SA. • While the tables very similar scores when excluding experts or items, we appreciate that the 1000-year discharge estimates for the decision maker might change significantly if more weight shifts from one to another expert.

Excluded expert	Information score total	Information score real.	Statistical accuracy
None	0.4852	0.4892	0.6828
Exp A	0.4338	0.4892	0.6828
Exp B	0.4852	0.4892	0.6828
Exp C	0.4389	0.418	0.6828
Exp D	0.4356	0.3913	0.7071
Exp E	0.4848	0.4892	0.6828
Exp F	0.4852	0.4892	0.6828
Exp G	0.4611	0.4761	0.6828

Excluded item	Information score total	Information score real.	Statistical accuracy
None	0.4852	0.4892	0.6828
ChaudfontaineT10	0.4924	0.5095	0.7059
ChoozT10	0.4841	0.4866	0.7059
GendronT10	0.4885	0.4988	0.7059
MartinriveT10	0.4854	0.4901	0.7059

SalzinnesT10	0.4833	0.4844	0.7059
TabreuxT10	0.4904	0.5039	0.3219
MembreT10	0.4857	0.4909	0.7059
StahT10	0.4706	0.449	0.7059
MeerssenT10	0.4821	0.481	0.5927
GochT10	0.4882	0.4979	0.4048

We did not include this robustness analysis in the main article, as it draws the focus too much to the details of the expert-elicitation. However, if the reviewer would find it suitable, we could include it as an appendix.

what if an expert's good judgment (i.e., closer to the true observation) was achieved by accident: Because experts are not assessed by the distance between, for example, the median and the realization but by their statistical accuracy (based on the number of answers to the calibration variables falling in each interquantile interval), the results should be relatively sensitive to a value (such as the median) that is accidentally close to the realization. Note also that the statistical accuracy is based on the total of estimates. For illustration, consider Figure 4. As long as an answer does not change interquantile interval, the SA will not change.

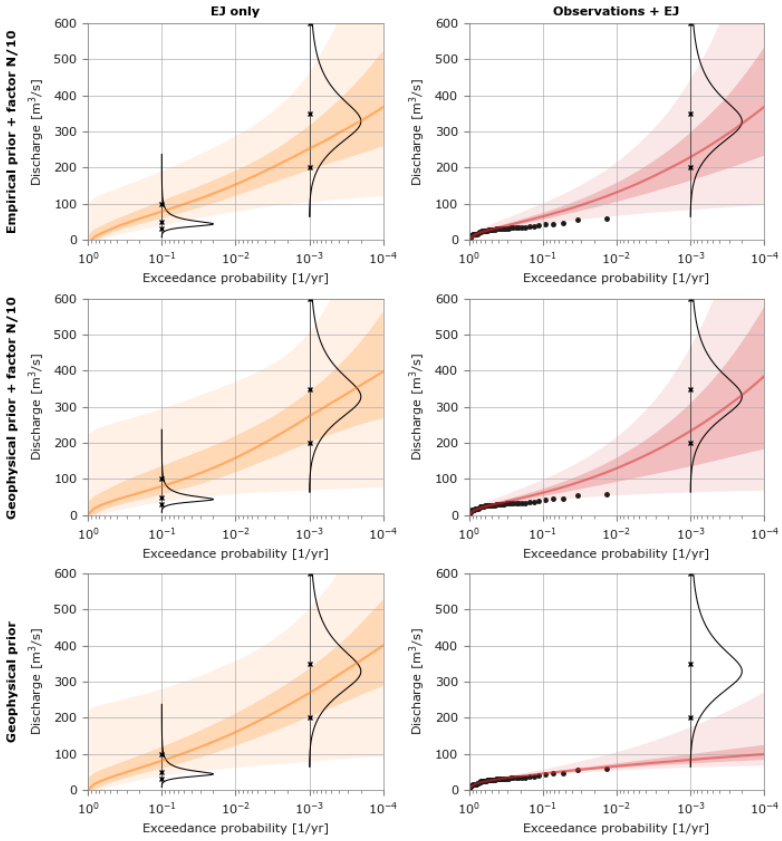
and his/her assumptions no longer work for a future event where the conditions have changed? Indeed Cooke's method relies heavily on the assumption that good statistical accuracy for seed variables (10-year return discharge estimates) is also a good statistical accuracy for variables of interest such as 1,000-year return discharges (or for future 10-year discharges, but this should be solved by applying results within the proper context). This is discussed in Section 5.2. "*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). This assumption is in fact one of the most crucial assumptions in Cooke's method and has extensively been discussed in for example Cooke (1991).*" (lines 487 to 490)

8A	8) I understand the Authors' reply and I am aware of the GRADE model. However, please understand that it is difficult to trust non-published material, when also, at the same time, hundreds of scientists struggle to find better and more accurate mathematical models to generate long-range rainfall and discharge timeseries. Also, it is clear from the results that the GRADE performed equally (if not better) than the experts' methods/models, and therefore, experts should base their judgment on this model to improve their own judgment. …	Regarding GRADE being non-published (at least the full method, the weather generator is): We appreciate this point. We added a note to this is the discussion (Section 5.2): "Finally, note that the full GRADE-method is not published in a peer-reviewed journal (the weather generator is, (Leander et al., 2005)). However, because the results are widely used in the Dutch practice of flood risk assessment (and known to the experts as well) we considered them the most suitable source for comparing the results in the present study." (lines 509 to 512)

8B	… Also, it is clear from the results that the GRADE performed equally (if not better) than the experts' methods/models, and therefore, experts should base their judgment on this model to improve their own judgment.	We did not explicitly handed the individual experts the GRADE statistics (or any other discharge statistics), in order to: 1) avoid influencing their estimates and 2) being able to compare the results to GRADE. As the expert study took place just after a 'new extreme', one may speculate that the experts' results would probably have been "GRADE plus some factor" if they knew the GRADE numbers. The study focuses on evaluating the method, in a case study to the Meuse River. We fully agree that the GRADE results could help improve the experts' estimates in case the study goal was to derive new Meuse EV-statistics. However, for our purpose, we did not present them with these data.

RESPONSE TO REFEREE 1'S COMMENTS (presented in report 2)		
#	**Referee comment**	**Authors' response**
1	…However, I am still skeptical about the fact that the ability to guess the magnitude of small floods implies the ability to guess the magnitude of large ones. I know that the experiment cannot be changed now but I am not satisfied by the discussion of the alternative. The Authors motivate their choice in Section 5.1 by saying that the 10-yr flood is a better target than model parameters because it is "observed". I would say that, as a quantile of the model/distribution, it is not observed, it is a model characteristic such as the moments or parameters. I find the "defense" of the choice in Section 5.1 rather weak. I would suggest indicating that an alternative choice could have been taken and, for example, could be tested in future work. The alternative choice (e.g., Renard et al., 2006, doi:10.1007/s00477-006-0047-4) is possible and, I would say, preferable. I strongly suggest that the Authors read Renard et al. (2006), as suggested in my first review, and discuss that alternative method in this paper…	We agree with the reviewer that the 10-yr flood is a quantile (a model characteristic) and not necessarily observed. Discharge is however a quantity that is used by hydrologists in their everyday work, and does not require a transformation from shape or scale parameter to discharge by the expert In principle, it is a measurable quantity (m^3/s for example). The approach presented in Renard et al. 2006 is indeed an alternative choice of which the authors were not aware when designing the study. We explain the main differences between Renard et al. and our study: 1. Renard et al. 2006 combine different models in their Bayesian estimation: Different distributions are combined, using a stationary, step, or sloped model for the location parameter in time. The authors acknowledge the merit of this method and therefore discuss it in more detail in the discussion. For this specific study, we did not adopt the 'multi-model' approach: • In terms of a varying location parameter in time: this is outside the scope of our study. • In terms of distributions, we use the GEV because we selected block maxima (and not peaks over threshold). The Gumbel is a particular case of the GEV-distribution, so we are satisfied with using just the GEV distribution and fitting the (possible Gumbel-) tail to the data. 2. When more than one quantile is used, in Renard et al. 2006 the difference between quantiles is used for any quantile after the first, instead of the quantile itself. This should reduce the dependence between the quantiles and therefore the priors as well. While no proof is provided by Coles and Tawn (1996) or Renard et al. (2006) that the difference in quantiles exhibit less dependence than the quantiles themselves this seems a reasonable assumption if the 1000-yr estimate is considered to be the sum of the 10-yr estimate and the estimate for the difference. In our study, two options are considered: • The combination of observed maxima and the EJ for the 1,000-year discharge. In this case only on EJ-quantile is used, so the approach is the same. • Using no observed maxima, but both the 10-year and 1,000-year discharges. In this case using the difference between quantiles would make a difference. Eliciting the differences would however come at the cost of the experts not being able to express their beliefs of the 1,000-year discharge directly in their estimate (which we find important, as explained before). 3. The use of a Jacobian to transform the prior, this is discussed at the last item. Note that the discussion Sect. 5.1 is adjusted to more properly explain the differences between Renard et al., 2006 and the present study. This mainly concerns the theoretical and practical differences between estimating discharges and differences between discharges.

2	… Regarding the Bayesian method, the Authors have made two major changes, i.e., using a reasonable prior for the GEV shape parameter, and removing the ad-hoc "weighting" procedure used in the first version of the paper. This is good. However, the language used should be corrected. I've never heard of prior likehood or posterior likehood in Bayesian statistics. I don't think the wording exists, please use proper wording (see e.g., https://en.wikipedia.org/wiki/Bayesian_inference#Formal_description_of_Bayesian_inference or any other Bayesian basics reference). …	We have updated the wording to correspond to formal use, as referred to by the reviewer. Additionally, we restructured Sect. 3.3 to put more focus on what information is used to fill the different parts of Bayes Theorem (prior distribution, likelihood), and how MCMC facilitates estimating the posterior distribution.

3	… Besides, since the equation on page 19 of the track-change document (no equation and line numbers there) differs from Eq. (6) in the original paper (the 10/N_i is no more in there), how comes that the results do not vary significantly? I would have liked to have an explanation in the reply to the reviewers (not in the new manuscript, of course). …	The individual GEV-fits do vary because of the change in method, which is most easily observed by comparing the results in Ch. 3 of the supplementary information in the initial manuscript to those in the supp. Information from the last revision (in which case it is Ch. 4). Compare for example the results for expert C, or expert D for Niers, Goch. The final results from the EJ decision makers do not vary significantly, because these are based on a weighted combination of individual experts. In some cases, the individual results go up, in others they go down. Moreover, the largest differences are for tributaries that do not contribute to the discharge at Borgharen (as the confluences are downstream of that location). The downstream results at Borgharen (location of interest in this study) are presented in the supplementary material Sect. 3.2 (initial submission) and Sect. 4.2 (last revision). Here we can see the uncertainties have become wider in the new results, but the medians for GL and EQ are largely unchanged. To illustrate the effects of changing: 1) The old prior to the geophysical prior 2) Removing the factor 10/N We show the intermediate step of only changing the prior in the following fit for expert D (high weight), and tributary Geul, Meerssen (downstream of Borgharen):  Note the labels on the left, which show which model choices are related to each plot. Left (orange) shows expert judgment only results, right (red) for expert judgment + observations. Changing the prior had a relatively limited effect (compare first and second row). Removing the factor 10/N (limiting the

		observations' weight in the posterior distribution, compare second and third row) had no effect in the EJ only results (since no observations), and a big effect on the combined results.
4	… Also, since the expert judgment is considered as a prior now, as the Authors claim, the equation on page 19 of the track-change document should express it in terms of model parameters and therefore I would have expected a Jacobian in front of g(F^-1(1-p|theta)) (see e.g. http://mystatisticsblog.blogspot.com/2018/04/jacobian-transformation-and-uniform.html).	The prior should indeed be expressed as a model parameter term (i.e., $\pi(\theta)$) but the experts are not estimating θ (or part of the vector) directly. As far as we can see, The effect of using the Jacobian in Renard et al., 2006 is clearest in the stationary exponential model, in which the Jacobian is the (partial) derivative of the quantile function to λ: $-\log(1-p)$. This corrects for the fact that the quantile function has a different derivative at different non-exceedance probabilities. In our view, such a transformation would be needed when using for example a non-informed (flat) prior, such as in the example in the link provided by the reviewer, such that a uniform estimate for the quantile would result in a uniform θ. For a non-exceedance probability p=0.999 this would be a larger 'factor' than for p=0.9. However, we prefer the prior $\pi(\theta)$ to follow the expert distribution g regardless of the elicited exceedance probability p. In summary, we did not change the method based on the Renard et al., 2006 article referred to by the reviewer, mainly because 1) the quantiles were elicited and not the difference between quantiles, 2) non-stationarity was out of scope, and 3) we do not think the use of a Jacobian is needed to transform the expert elicited probability density. We do however acknowledge the merits of the approach and have given it a proper discussion in 5.1, such that readers of the article will not be unaware of the approach.