The authors would like to thank the reviewer for reviewing the article. We have written a response to the reviewer's comments below. If the reviewer agrees with our interpretation of the comments and the responses, we will process this as textual changes in the final version of the article (depending on how that proceeds).

This study investigates through the Cooke's method how scientific judgments by experts can assist flood-risk managers. In my opinion, there are several issues that need to be addressed so that the applied methods, justification, and results, can be clearer and of practical use to other case studies. Please see several such comments and suggestions below:

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.

That is a good suggestion. We'll rephrase parts of the introduction such that it doesn't seem like a "physics-based modelling does not incorporate uncertainty" statement. Right now it does not rightly acknowledge the variety in different modelling approach, and their respective assessment of uncertainty (from which Dimitriadis et al. 2016) is an example).

Dimitriadis, P., A. Tegos, A. Oikonomou, V. Pagana, A. Koukouvinos, N. Mamassis, D. Koutsoyiannis, and A. Efstratiadis, Comparative evaluation of 1D and quasi-2D hydraulic models based on benchmark and real-world applications for uncertainty assessment in flood mapping, Journal of Hydrology, 534, 478–492, doi:10.1016/j.jhydrol.2016.01.020, 2016.

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.

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 (hydro-power as well as weirs). The main cause of the event being surprisingly large was however the meteorological situation. The most important point here is that practitioners need to be aware of the uncertainties in their modelling approach. We'll try and add this to the introductory text.

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).

This is a good point. The current wording indeed suggest that historic data carry a deterministic return period that can be extrapolated like a data point, while this requires modelling assumptions (such as a probability distribution). We will rephrase this.

Koutsoyiannis, D., Replacing histogram with smooth empirical probability density function estimated by K-moments, Sci, 4 (4), 50, doi:10.3390/sci4040050, 2022.

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 model (Eq. 1), is a framework to process the experts' estimates. The experts had to come up with 10-year and 1000-year discharge estimates (5^{th}, 50^{th} and 95^{th} percentiles, such that it is an uncertainty assessment). The experts were free in choosing their methods. It was however a 1-day expert session, 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 themselves. The model framework was discussed with them, and they had to fill in the numbers with which the estimates of extremes can be generated (through the model framework).

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 will clarify Section 3.4. It suggests that we used 10.000 samples for each tributary, but this process is repeated 2000 times. The 10.000 realizations for each tributary are used to generate an exceedance frequency curve for the downstream location. By doing this 2000 times, we also get uncertainty bounds for this. These 10.000 realizations represent 10.000 years, which is sufficient for estimating a 1000-year flow, given that the process is repeated 2000 times.

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 will add some numbers from 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 610-year average recurrence interval when considering only the summer half year (April to September). These estimates are based on MCMC-fitted GEV-distributions including the 2021 event. Please refer to figure 2.5 in that report (unfortunately Dutch only).

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.

Agree, we'll add a comment that, while often not explicitly, all modelling involves (or at least should) expert judgment to some extent.

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 will add some more details to the article, but the reviewer's guess is right: It resamples the historically observed rainfall while preserving the spatial and temporal correlation. This rainfall is then processed through a hydrologic model to generate tributary discharges that are simulated with a hydraulic model.