[General reply from the authors]

We would like to thank the reviewer for taking the time to review our manuscript. We highly appreciate her suggestions and comments, which are helpful in improving the manuscript. Below we have replied to the various comments made by the reviewer.

[Replies to reviewer comments]

1. The Authors introduce the bootstrap approach (l. 5-9 p. 3) as a solution to overcome the problem of isolated historical events for which confidence intervals are typically not symmetrical. It is not clear what the Authors mean by symmetrical confidence intervals; this issue should be explained since it is the motivation (together with the easy application of FFA) for reconstructing a continuous data set.

Symmetrical means that the confidence intervals follow a normal distribution. Hence, the 95% confidence intervals can be computed with the basic rule of +/- 1.96*standard deviation. However, confidence intervals are typically not symmetrical for flood frequency relations. Hence, these intervals are difficult to compute if the data of annual maximum discharges is extended with historic events in isolation. Therefore, we would like to create a continuous data set such that the method to compute confidence intervals remains unchanged compared to traditional FFAs. Please also see the next comment.

Further, bootstrap is not necessary for confidence interval estimation (l. 9-10 p. 3) yet still necessary for continuous data set reconstruction.

It is indeed true that a bootstrap approach is not needed to compute the confidence intervals if a continuous data set is present. However, a bootstrap method is still needed to create a continuous data set as was done in this study. Both are different kind of bootstrap approaches. Using the same terminology leads to confusion. Therefore, the section will be adapted to (with in green the new sentences):

"The objective is to develop a straightforward method to consider historic flood events in an FFA, while the basic principles of an FFA remain unchanged. Confidence intervals of flood frequency relations are typically not symmetrical distributed (Schendel and Thongwichian, 2017). This means that the confidence intervals do not obey a normal distribution, but they are skewed. For a continuous data set, the asymmetric distributed confidence intervals can be computed relatively easily, while this becomes more problematic if historic flood events are added to the data set of measured discharges in isolation. To overcome this problem, bootstrap approaches such as the test inversion bootstrap method are recently developed (e.g. Burn (2003); Kyselý (2008); Schendel and Thongwichian (2017)). This study is novel since a continuous data set is created. The use of a bootstrap approach to compute the confidence intervals is now redundant. Although still a bootstrap approach is required to create the continuous data set, the method to compute the confidence intervals does not change compared to an FFA solely based on measured annual maximum discharges. This makes the comparison between the confidence intervals of the measured annual maximum discharges and the extended data set more reasonable."

2. The hydraulic model is used to propagate the discharge for the historic flood events reconstructed by Meurs (2006) from Cologne to Lobith; to this aim the Authors state that they use the current geometry of the riverbed and floodplain in order to correct the historic floods for anthropogenic interventions and natural changes of the river system, which is referred as "normalization" in the manuscript (l. 10-14 p.3). This approach is unusual based on my experience (Calenda et al., 2005); historical flood events should be simulated by reconstructing the historical conditions (the river geometry as in the period the flood occur), that is what Authors would have available if measures would have started in the ancient past. In essence, I am not convinced that propagating the ancient floods in the current riverbed is the correct approach to solve the "homogenization" problem; conversely, this "gives insight in the consequences of an event with the same characteristics of a historic flood event translated to present times" (as stated by the Authors themselves at l. 17-18, p. 3).

It is indeed true that historic flood events should be reconstructed based on the historical conditions. This is exactly what Meurs (2006) has done. Historic flood events were reconstructed near the city of Cologne, Germany, based on reconstructed main channel bathymetry.

However, our aim in this paper was not to make reconstructions of the historic events along the river stretch. In this paper, we aimed to predict flood frequency relations for current water policy assessments and therefore we would like to have the present-day discharges. This is why 'normalization' is done in the Dutch water policy. Even the measured discharges in e.g. 1920 are normalized to present-day discharges since the river system has altered a lot due to human interventions resulting in a change of the flood frequency relation. Nowadays, more water is capable of flowing through the river system towards Lobith, German-Dutch border, as a result of the heightened dikes along the Lower Rhine. Therefore, the historic flood events have no predictive value without normalzing it into present-day discharges. This is why we have normalized the historic flood events at Cologne, which are based on historical information, to present-day discharges at Lobith. To do so, we use the hydraulic model which is based on the current geometry. This hydraulic model is described in Bomers et al. (2019), and now accepted for publication in Natural Hazards.

3. Based on my opinion the Authors should "naturalize" the estimated discharge, by computing the discharge that they would have observed in absence of some anthropogenic change in the riverbed or in the catchment (l. 14-16 p.3). This means that are the recent events that should be reported to pre-dike conditions and not the opposite (as done in Section 2.3.2). The presence of the dike artificially alters the natural regime of the extreme flood events; the anthropogenic alteration of flood regime should be of deterministic nature, even if its estimation is characterized by a certain degree of uncertainty.

For flood safety assessments, we are interested in the current flooding regime and not that of the pre-dike conditions. It is indeed true that the presence of the dike alters the natural regime of the extreme flood events, but we are interested in this change since it determines how much water can enter the Netherlands at Lobith nowadays. Therefore, normalization of the historic flood events to present-day conditions is of high importance to correctly estimate flood frequency relations of the present river system.

4. Why do the normalized events almost always lead to a higher discharge than the historic event (l. 16-17, p. 3)?

This is because more water is capable of flowing through the river system as a result of the heightened dikes along the Lower Rhine. Nowadays, floods occur for higher discharge stages compared to the historical time period. This will be added in the revised manuscript

5. Section 2. For the sake of clarity, a table summarizing the type of information and the related uncertainty for the different time periods should be included.

A table with the various types of uncertainties for each time period will be added in the revised manuscript. See table below.

Time period	Data source	Cause uncertainty	Location
1342-1771	Meurs (2006)	Reconstruction uncertainty caused by uncertain main channel bathymetry, bed friction and maximum occurred water levels	Andernach
1772-1865	Toonen (2015)	Reconstruction uncertainty based on measured water levels of surrounding sites	Emmerich, Pannerden and Nijmegen
1866-1900	Toonen (2015)	Uncertainty caused by translation measured water levels into discharges	Lobith
1901-1950	Tijssen (2008)	Uncertainty caused by extrapolation techniques to translate measured velocities at the water surface into discharges	Lobith
1950-2000	Tijssen (2008)	Uncertainty caused by translation velocity-depth profiles into discharges	Lobith
2000-2008	Tijssen (2008)	Measurement errors for discharges slightly exceeding the bankfull discharge	Lobith
2008-2018	Measured water levels available at https://waterinfo.rws.nl	Measurement errors for discharges slightly exceeding the bankfull discharge	Lobith

6. L. 14-15, p. 4. The Authors should clarify the distance and the characteristics of the nearby gauging locations.

The following will be added in the manuscript:

"For the period 1772-1865 water levels were measured at the nearby gauging locations Emmerich (Germany) located 10 kilometers in upstream direction, Pannerden located 10 kilometers in downstream direction and Nijmegen located 22 kilometers in downstream direction."

However, note that this analysis has been performed by Toonen (2015) and is not part of this paper. Therefore, we refer for more information about the characteristics of the 1772-1901 data set to Toonen (2015).

7. The procedure discussed in Section 3 is based on a non-parametric approach; alternatively a parametric method, based on the same assumption that ancient flood events follow the same statistical behavior of those systematically recorded, could have been considered. See Stedinger and Cohn (1986) and Francés

It is indeed true that a non-parametric approach could have been considered. However, in this paper we had the preference to create a continuous data set instead. This is because, since recently, the Dutch water policy uses a new method in which a continuous data set of 50,000 years based on resampled measured weather conditions (e.g. rainfall, temperature, evapotranspiration) is used to predict flood frequency relations (Hegnauer et al., 2014, and also described in Chbab (2006)). We wanted to use the method of Hegnauer et al. (2014) of creating a continuous data set to test whether it also works with resampling measured discharges. This makes the use of HBV and Hydraulic modelling to translate the weather data into maximum discharges redundant, as was done by Hegnauer et al. (2014).

Furthermore, we wanted to create a continuous data set since the computation of the confidence intervals of a flood frequency relation remains unchanged compared to the analysis of just measured annual maximum discharges, making the comparison between the two more reasonable. This argument will be added in the introduction of the revised manuscript. For future work, it is interesting to study how confidence intervals deviate between the proposed methodology and a method based on a parametric approach. However, our results are in line with the findings of Francés (1998), who also showed that the uncertainty intervals of FFAs reduces if historical information is included in the analysis.

8. L. 2-7, p. 9. The Authors states that "the available goodness-of-fit tests for selecting an appropriate distribution function are often inconclusive. Those tests are more appropriate for the central part of the distribution than for the tail (Chbab et al., 2006), where we are interested in since the tail determines the investments required for future flood protection measures." I agree with the Authors that goodness-of-fit tests might be inconclusive, as discussed deeply in Serinaldi et al. (2018); on the other hand they provide a first indication on which models, among several competing ones, could be excluded due to the poor performance (see, e.g., Laio, 2004). In such a sense, I suggest the Authors at least to rephrase the sentence, also because there are different goodness-of-fit test which focus on the statistical behavior of the tails, such as the Anderson-Darling test and the Modified Anderson-darling test (Laio, 2004).

We agree with you that there are various goodness-of-fit tests, all with their own properties. The sentence will be rewritten with in green the new text:

"A probability distribution function is used to fit the annual maximum discharges to its probability of occurrence. Many types of distribution functions and goodness-of-fit tests exist, all with their own properties and drawbacks. However, the available goodness-of-fit tests for selecting an appropriate distribution function are often inconclusive. This is mainly because each test is more appropriate for a specific part of the distribution, while we are interested in the overall fit of the distribution. This is because the safety standards expressed in probability of flooding along the Dutch dikes vary from 10^{-2} to 10^{-5}. We restrict our analysis to the use of a Generalized Extreme Value (GEV) distribution since this is commonly used in literature to perform an FFA"

9. Following the argument of previous comment, I do not believe that restricting the analysis to a single probability distribution model (although it is the Generalized Extreme Value distribution commonly used in literature to perform an FFA) is a good choice. Since the interest is in evaluating how the confidence bounds of extreme quantile estimates reduce when adding the historical information (l. 18- 21 p. 9), it should be considered that confidence bounds depend not only on the length and information content of the dataset but also on the probability model itself. Hence, results could be different if a different model is taken into account.

You are indeed correct that the uncertainty interval also highly depends on the fitted distribution itself. Although not shown, we performed the analysis with other distributions as well (e.g. Weibull and Gumbel) and the general conclusion of 'reduction of the confidence bounds as a result of extending the data set of measured discharges' also holds for these distributions. For the GEV distribution we found a reduction of 73% as a result of extending the data set of annual measured discharges with historic events, with the Gumbel distribution a reduction of 60% and with the Weibull distribution a reduction of 67%.

We will add a sentence to the revised manuscript in which it is stated that also for other distribution functions a reduction of the confidence interval was found. However, we will not show the in-depth results of different distribution types, because we think this is distracting the reader from the analysis performed and corresponding main findings. Furthermore, the GEV distribution has been shown to fit the data of the Rhine river well and therefore this distribution was preferred above other distributions. Finally, we would like to highlight that many closely-related studies also only focused on the use of a single distribution (e.g. Francés (1998)).

10. L. 10-12 p. 9. Do you the Authors mean that they assume an upper bounded distribution? This issue should be clarified.

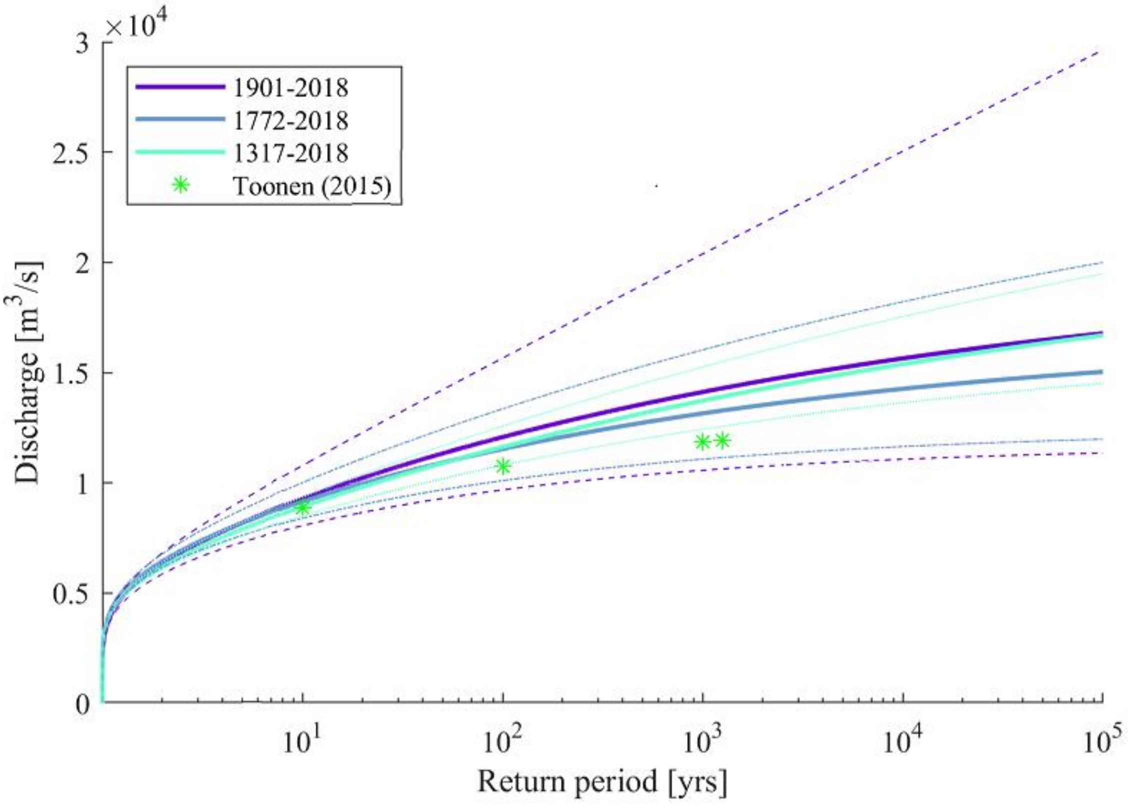
Yes, we indeed assume an upper bounded distribution. The GEV distribution has an upper bound as a result of the shape parameter which both influences the skewness and kurtosis of the distribution. We use a bounded distribution since the maximum discharge that is capable of entering the Netherlands at Lobith is limited to a physical maximum value. The crest levels of the dikes along the Lower Rhine are not infinitely high. The height of the dikes influences the discharge capacity of the Lower Rhine and hence the discharge that can flow towards Lobith. This explanation will be added to the revised manuscript such that it becomes clear why we use an upper bounded distribution.

11. Figure 5 is unnecessary, It could be removed.

Figure 5 will be removed from the revised manuscript.

12. Figure 6. The largest extreme events are not included in the uncertainty bounds. The corresponding sample bounds could be included as well to text the model performance (see comment 9).

Also the largest extreme events are included in the uncertainty bounds (see table 1). However, since the upper bound of the measured data set has a value of 29,631 m3/s (table 1) this line was not entirely drawn. Since it leads to confusion, we will plot the entire line in the revised manuscript. See the figure below.



13. Section 5.2. I am not sure I fully understood the rationale and the approach behind the analysis performed here. The historical events are some of the highest events observed in the whole observation period. If a sample is reconstructed by simply resampling the events observed in 1901-2018 (without including the largest historical events but with the same length of that used in previous sections), the largest events might only be those observed in the more recent period; as a consequence, the fitted model is expected to be characterized by, e.g., a smaller variance, which implies narrower uncertainty bounds. I do not see this behavior in figure 7 (upper panel). What I see in figure 7 is that the fitted model in the two cases is almost the same, while the uncertainty bounds are significantly different. I can explain this only if the reconstructed samples have a very different length. Please provide a deeper explanation.

You are indeed correct that we simply resample the events observed in 1901-2018 without including the largest historical events but with the same length. This corresponds with the line 'Bootstrap 1901-2018 data'. This data set has a length equal to the 1317-2018 period. If we compare the line with the Bootstrap 1317-2018 data set, we indeed see that the uncertainty interval of the Bootstrap 1901-2018 is still larger even though the length of the two data sets are the same. It must be noted that not only the length influences the uncertainty interval, but also the discharges within the data set and resulting variance.

For the Bootstrap 1901-2018 data set, the entire measured data set (1901-2018) is used for resampling. The created continuous series (5,000 in total for convergence reasons) has an average variance of 4,19 x 10^6 m^3/s. For the Bootstrap 1772-2018 data set, only the discharges below a certain threshold in the measured time period (1772-2018) are used for resampling. In this study, the perception threshold was chosen to be equal to the lowest flood event in the historical time period having a discharge of between 6,928-10,724 m^3/s. Hence, the missing years in the historical time period are filled with relatively low discharges, but some of the largest events in the historical time period are larger than ever measured. The total variance of the data set decreases (3.35 x 10^6 m^3/s) as a result of the lower discharges to create the continuous data set. As a result of the lower variance, also the uncertainty bounds are smaller compared to the Bootstrap 1901-2018 data set. This explanation will be added in the revised manuscript.

14. L. 20-22 p. 14. It is not clear how the extended data set with normalized reconstructed discharges can capture the long-term climatic variability (see also previous comments).

The historic flood events are only normalized for changes in the river system. As a result, the normalized discharges still capture the climatic conditions in the historical time period. Although the missing years within the historical time period are filled with the measured data set 1772-2018, the most extreme events still capture the climatic variability in the period ~1300-2018. This will be added in the revised manuscript.

15. L. 35, p. 14. Isn't it the 1374 event?

The 1374 flood event is indeed the largest observed discharge (at Cologne) of the last 1,000 years. However, in this analysis we consider the largest measured discharge (measurements have been performed since 1900), which correspond with the 1926 flood event.

16. Fig. 8. Adding one event equal to the largest one over a record is expected to affect somewhat the estimated model if the record is 100 years while non changes in the model are expected if the record is about 700 years. Hence, which is the lesson learned from this analysis?

The lesson learned is that flood safety assessments become more robust if the data set of annual maximum discharges is extended. After the 1993 and 1995 flood events of the Rhine river, the flood frequency relation altered significantly resulting in an increase of the design discharge at Lobith of 1,000 m3/s. Such an increase in the design discharge requires huge investments to cope with the new flood safety standards which were set after the 1993 and 1995 floods. Such an increase was not found if a longer time series was included in the analysis. Looking at the results, decision makers might have taken a different decision.

17. Within the Conclusion Section a detailed list of the limitations of the approach proposed here should be provided.

A list of the limitations of the proposed method will be included in the discussion which is:

- The 1772-2018 measured data set is used to create a continuous data set. Preferably, we would like to have a historical continuous record since now the low flows (discharges with high probability of occurrence) are biased on climate conditions of the last 250 years
- Historical flood events must be normalized for anthropogenic and natural changes in the river system which can be quite time demanding in terms of computational time
- The extended data set still has limited length. Preferable we would like to have a data set of e.g. 100,000 years such that extrapolation to such return periods is not required anymore. However, the proposed method is a large step to decrease uncertainty.

- The predicted uncertainty intervals depend on the chosen perception threshold. A larger threshold results in an increase of the variance of the data set and hence to an increase in the uncertainty intervals.
- The shape of the constructed FF curve strongly depends on the climate conditions of the period considered. If the data set is extended with a period which only has a small number of large flood events, this will result in a significant shift of the FF curve in downward direction. This shift can be overestimated if the absence of large flood events only applies to the period used to extend the data set.

Up till now, only the last point was mentioned in the discussion. The other points will be added.

References

Bomers, A., Schielen, R.M.J., Hulscher, S.J.M.H., 2019. Consequences of dike breaches and dike overflow in a bifurcating river system. Accepted for publication in: *Natural Hazards.*

Hegnauer, M., Beersma, J.J., van den Boogaard, H.F.P., Buishand, T.A., Passchier, R.H., 2014.Generator of Rainfall and Discharge Extremes (GRADE) for the Rhine and Meuse basins. Final report of GRADE 2.0. Technical Report. Deltares. Delft, The Netherlands