[General reply from the authors]

We would like to thank the anonymous reviewer for taking the time to review our manuscript. We highly appreciate the 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]

This paper discuss the extension of a flood series, based on hydraulic modelling and the utilisation of extend hydrological time series to estimate the frequency of extreme floods at the Rhine gauge Lobith. The authors expect a reduced sampling effect. It is widely known that for extreme events the empirical exceedance probabilities in short observation series are often overestimated. To solve this problem the authors suggest to extend the observed time series. In their case study they propose to extent the existing series of observations between 1901-2018 by a linear regression of water levels with neighbouring gauges for the period 1772 to 1900 based on a previous study from Toonen (Toonen, 2015) and the translation of these water levels into discharges using a stage-discharge relationship, which is not specified in detail.

It is indeed true that the translation of the water levels into discharges for the period 1772-1900 was not specified in detail. This is because this has been described in detail by Toonen (2015) who performed the analysis. To help the readers of our manuscript, the following will be added in the revised manuscript in section 2.2, with in green the new text:

"For the period 1772-1900, the data presented by Toonen (2015) is used. At Lobith, daily water level measurements are available since 1866. For the period 1772-1865 water levels were measured at the nearby gauging locations Emmerich, Pannerden and Nijmegen. Toonen (2015) used the water levels of these locations to compute the water level at Lobith and associated uncertainty interval with the use of linear regression between the different measurement locations. Subsequently, he translated these water levels, together with the measured water levels for the period 1866-1900, into discharges using stage-discharge relations at Lobith. These relations were derived based on discharge predictions adopted from Cologne before 1900 and measured discharges at Lobith after 1900, and water levels estimates from the measurement locations Emmerich, Pannerden, Nijmegen and Lobith. Since the discharge at Cologne strongly correlates with the discharge at Lobith, the measured discharges in the period 1817-1900 could be used to predict discharges at Lobith. Hence, the reconstructed water levels were used to derive stage-discharge relations. The 95% confidence interval in reconstructed water levels propagates in the application of stage-discharge relations, resulting in an uncertainty range of approximately 12% for the reconstructed discharges (Fig. 1). The reconstructed discharges in the period 1772-1900 represent the computed maximum discharges at the time of occurrence and has not been normalized for changes in the river system."

The resulting series (1772-2018) is named as the "systematic" time period. The other and even more uncertain step consists in an estimation of the peaks of historic floods at Lobith. Here a series of 12 historic flood events in Cologne since 1342, provided by Meures and Herget is used. As these events were estimated more than 150 km upstreams, a (1D-2D) coupled hydraulic model is used to transfer these peaks to Lobith: "The reconstructed maximum discharges at Cologne (Meurs, 2006), which are not normalized for anthropogenic interventions upstream of Cologne, are used to predict maximum discharges at Lobith with the use of a hydraulic model to normalize the data set." The meaning of

"normalization" in this context stays unclear. It seems to be the adaptation of these peaks (which were roughly estimated by Meures) on to today's conditions.

We indeed mean with normalization adapting the historic peaks at Cologne on today's geometry conditions. Hence we will find the maximum discharges at Lobith as a result of the maximum discharges at Cologne under current river conditions. Please see page 3, lines 13-14: "In such a way, the historic floods are corrected for anthropogenic interventions and natural changes of the river system, referred to as normalization in this study."

There are extreme uncertainties connected with this approach: the river reach changed in its hydraulic characteristics over 700 years, the water levels in Cologne dating back several hundreds of years are uncertain, the discharges as well and so on. It is a big surprise that the authors are able to specify in Fig. 3 95% confidence intervals for the maximum discharges in Cologne and Lobith for these 12 events. It stays unclear how these intervals were estimated.

The 95% confidence interval for the maximum discharges in Cologne were taken from Meurs (2006). His method is shown by Herget and Meurs (2010) in detail, using the 1374 flood event as a case study. The following will be added in the revised manuscript in section 2.3 with in green the new text:

"Meurs (2006) has reconstructed maximum discharges during historic flood events near the city of Cologne (Germany). The oldest event dates back to 1342. The used method is described in detail by Herget and Meurs (2010), in which the 1374 flood event was used as a case study. Historic documents providing information about the maximum water level during the flood event were combined with the reconstruction of the river cross section at that same time. Herget and Meurs (2010) calculated mean flow velocities near the city of Cologne at the time of the historic flood events with the use of the empirical Manning's equation:

$$Q_p = A_p {R_p}^{2/3} S^{1/2} n^{-1}$$

where Q_p represents the peak discharge, A_p the cross-sectional area during the highest flood level, R_p the hydraulic radius during the highest flood level, S the slope and n the Manning's roughness coefficient.

However, the highest flood level as well as Manning's roughness coefficient are uncertain. The range of maximum water levels was based on historical sources, whereas the range of Manning's roughness coefficients were based on the tables of Chow (1959). With this information, Herget and Meurs (2010) were able to calculate maximum discharges of the specific historic flood events and associated uncertainty range (Fig. 3)."

The reconstructed historic discharges and their uncertainties were used as input data of the 1D-2D coupled model to compute resulting discharges at Lobith. This is a valid method since there is a strong correlation between the discharge at Cologne and Lobith for in channel flow conditions, even though Cologne is located roughly 160 km upstream of Lobith since they are located in the same fluvial trunk valley and only have minor tributaries (Sieg, Ruhr and Lippe) joining in between (Toonen, 2015). This will be added in the revised manuscript to clarify the applicability of using historical discharge reconstructions at Cologne to determine corresponding present-day maximum discharges at Lobith.

With the 1D-2D coupled model, a Monte Carlo analysis was performed for each historic flood event in which the following parameters were considered to be random: maximum upstream discharge (based on the uncertainty range of each historic flood event as reconstructed by Herget and Meurs (2010)), dike breach thresholds, dike breach formation time and final breach width. The method of this analysis is discussed in detail by Bomers et al. (2019) and has recently been accepted for publication. As a result of the uncertain upstream discharge and breach characteristics, also the discharge at Lobith for each historic event is uncertain. Therefore, many model runs are performed for each event until convergence in model results is reached. Hence, the expected discharge at Lobith and expected 95% confidence intervals were computed (Fig. 3). The hydraulic modelling approach to normalize the historic flood events will be explained in more detail in the revised manuscript.

The authors propose a bootstrap sampling method to fill the gaps between the historic floods with annual flood peaks from the systematic data set, that have an expected value lower than the sampled perception threshold which is set as the smallest flood among the historic peaks. This is approach seems to be critical as it does not add any information to the statistical analysis. The todays conditions are modified by the first extension to the part of the series until 1772. With the sampling the authors accept that the flood series consist of independent and identically distributed random variables, which is not certain.

Indeed, we assume independent and identically distributed random variables. The authors are aware of this assumption. However, please note that to perform a flood frequency analysis we always have to assume that the discharge observations are independent and stationary (Khaliq et al., 2006). Although the assumption is highly uncertain, it must be noted that up till now no consistent large-scale climate change signal in observed flood magnitudes has been identified (Blöschl et al., 2017) justifying the assumption of independent and identically distributed random variables.

By definition bootstrapping is any test or metric that relies on random sampling with replacement. Here the wording “resampling of the non-systematic time series below the perception threshold” would be more appropriated. This has been done 5000 times and also the historical floods are varied within their 95% confidence intervals (however these were estimated!). The systematic series were not changed.

Maybe this was not fully clear to the reviewer, but also the systematic data set was changed. For each year within the historical period of which no data is available, an annual maximum discharge of the systematic data set below the perception threshold was randomly drawn (See step 5 in Fig. 4). This corresponds with the bootstrap method. As a result, each created continuous data set is different.

Furthermore, the values within the systematic data set were varied within their 95% confidence intervals. The study described the uncertainties of the systematic data set, which vary for different time periods as a result of different measurement methods used. Please see Fig. 1 in the manuscript and Section 2.1. We will add a table in which all types of uncertainties are described for the various data sets used to extend the data set of maximum discharges.

The GEV was estimated for each of these samples, the distributions were averaged (!) and their 95% percent confidence bounds were estimated. Table 1 specifies these 95% bounds with the 2-sigma-

reach, this would be only justified if the quantiles would be normal distributed. I suppose that this is not the case.

Indeed, the confidence bounds of the discharges are not normally distributed. The caption of the table stating 2-sigma is not correct. If you look at the numbers of the uncertainty bounds you can already see that the confidence bounds are not normal distributed since the upper bound is much further away from the average value than the lower bound, specifically for a return period of 100,000 years. The caption will be changed accordingly in the revised manuscript.

In total the value of this resampling study stays unclear for me as it does not extend the information content. The information, derived from the systematic series are used in a simulation study, but the basic assumption that the floods between 1772 and 1900 are reconstructed correctly adds uncertainty to it.

We indeed assume that the 1772-1900 flood were reconstructed correctly, but not without uncertainty. We have included this uncertainty in the analysis. The 95% bounds of the 1772-1900 data set are determined by Toonen (2015) and explained in more detail above. He found an uncertainty interval of approximately 12%. This will be added explicitly to the revised manuscript to avoid further misunderstanding. Furthermore, we will add a table describing the uncertainties of the various data sets used to extend the systematic data set.

There are at least two other options to consider historic floods in statistics:

REIS D. S., JR.; STEDINGER J. R. (2005): Bayesian MCMC flood frequency analysis with historical information. In: Journal of Hydrology, 313, pp. 97–116 (cited by the authors)

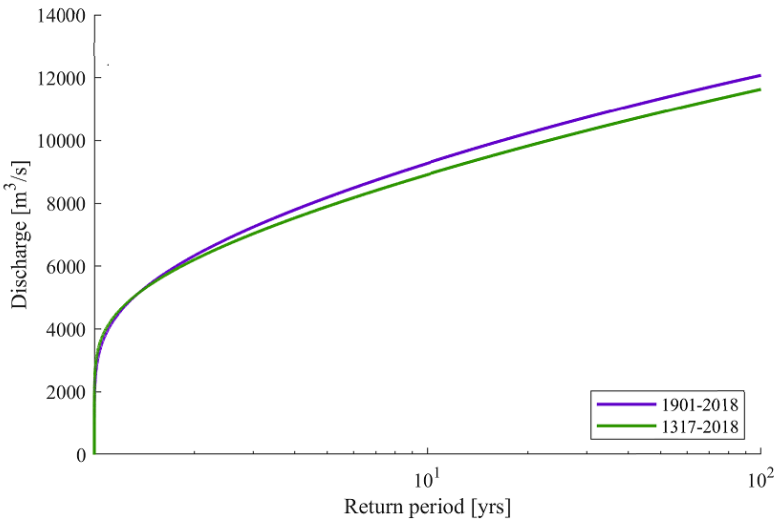
Wang, Q. J. (1990): Unbiased estimation of probability weighted moments and partial probability weighted moments from systematic and historical flood information and their application to estimating the GEV distribution. In: Journal of Hydrology 120 (1-4), S. 115–124

Both methods combine the information from the systematic data with historic floods without assumption that these observations are representative for the historic series. In both methods, it is assumed that the historic floods are representative for today's conditions. These events are used to improve the estimation of the upper tail only. The systematic part of the series stays untouched. In this way the uncertainty of assumptions of a large part of the time series is avoided. The statement of the authors: "Most studies found that the confidence intervals of design discharges were reduced significantly by extending the systematic data set with historic events." does not mean that an artificially extended systematic dataset would be beneficial if it was expanded with uncertain assumptions about past flood conditions and their adaptation to the current situation.

We agree with you that we add uncertainty to the data set by adding historical flood events to the measured data set and by using a resampling method to create a continuous data set. However, it must be noted that many of the uncertainties of the historic flood events are included in the analysis, as well as the uncertainty of the systematic data set (1772-2018). The 95% confidence intervals of the flood frequency relations are thus based on these uncertainties.

It is true that our method influences the flood frequency curve in the domain of the systematic data set (discharges with high probability of occurrence). However, as far as we know this is always the case if the parameters of the (GEV) distribution are recomputed as a result of new data availability. If

we have a look at the figure below, we find that the design discharge with a return period of 100 years decreases from 12,080 m^3/s to 11,630 m^3/s by extending the systematic 1901-2018 data set towards 1317 using the bootstrap method. This decrease in design discharge corresponds with a change of 3.7% indicating that resampling the systematic data set of the historical time period only has a little effect on the shape of the flood frequency curve corresponding with high probability of occurrence. This justifies the use of the bootstrap method. Furthermore, we would like to highlight that we are typically interested in correct prediction of the tail, rather than the discharges with large probability of occurrence, since the tail (high return periods) is of high importance to design flood protection measures.



My summary: The manuscript has some weakness with regard to uncertainty assessments (confidence intervals) where the methodology is not sufficient described. The assumption of a symmetrical interval seems to be arbitrarily. Nevertheless the topic is interesting, the manuscript should be consider the existing state of the art in this field and compare its results with well-established existing methods. I suggest to reject the manuscript for major revisions.

We agree with the reviewer that we did not provide enough details about the considered uncertainties of the various data sets used. We will provide a detail explanation of the computed 95% confidence intervals of the following data sets in the revised manuscript:

- Reconstructed historic flood events at Cologne by Meurs (2006)
- Corresponding historic discharges at Lobith using a hydraulic model
- Reconstructed discharges for the period 1772-1900 by Toonen (2015)
- Measured discharges for the period 1901-2018

Furthermore, we will add more information about why we propose this method instead of a Bayesian method. We would like to highlight that our method is systematic. We can extend our data set with historical data and keep the method of a flood frequency analysis the same. In this way, we can make a clear comparison on the effect of extending the data set with multiple other sets on the confidence bounds of flood frequency analysis.

Although the maximum likelihood method only gives a point estimate of the (GEV) parameters, as sample size increases, maximum likelihood estimators become unbiased minimum variance estimators with approximate normal distributions. This is used to compute confidence bounds for

the GEV parameter estimates. We would like to highlight that, although the Bayesian method is capable of predicting parameter uncertainty without the assumption of being normally distributed, the results are influenced by the prior. The influence of the prior, which has to be defined by the modeler, on the posterior distribution of the parameters and hence on the uncertainty of flood frequency relations can even be larger than the influence of discharge measurement errors, as was found by Neppel et al. (2010). The disadvantage is thus that we have to choose the prior in the Bayesian method correctly such that the tail will be correctly predicted. However, we do not have any measurements in, or near to, the tail and consequently it is reasonable to estimate the prior by fitting the original data with the use of e.g. the Maximum Likelihood method. In this way, the benefits of the Bayesian method compared to a traditional flood frequency analysis are at least questionable.

We are aware that there is a strong debate between the 'Bayesians' and the 'Frequentist' in literature and discussion forums. With this paper, we do not want to get into this discussion. Rather, we wanted to show a novel and systematic approach which is easy to understand for practitioners to include historic flood information into flood safety assessments. The general methodology of a flood frequency analysis remains in this proposed bootstrap methodology, only the data set of measured discharges is extended. As a result, this method is close to current practice of water managers. We will add the reasons why we set up a bootstrap method in the revised manuscript and will compare the methodology with the Bayesian statistics.

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. doi: 10.1007/s11069-019-03643-y.

Böschl, G., Hall, J., Parajka, J., Perdigão, R.A.P., Merz, B., et al. (2017) Changing climate shifts timing of European floods. In: Science 357, pp. 588–590. doi:10.1126/science.aan2506.

Frances, F. (1998) Using the TCEV distribution function with systematic and non-systematic data in a regional flood frequency analysis. In: Stochastic Hydrology and Hydraulics 12, pp. 267-283.

Khaliq, M.N., Ouarda, T.B., Ondo, J.C., Gachon, P., Bobée, B. (2006) Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. In: Journal of Hydrology 329, pp. 534–552. doi:10.1016/j.jhydrol.2006.03.004

Neppel, L., Renard, B., Lang, M., Ayral, P.a., Coeur, D., Gaume, E., Jacob, N., Payrastre, O., Pobanz, K., Vinet, F.,(2010) Flood frequency analysis using historical data: accounting for random and systematic errors. In: Hydrological Sciences Journal 55, pp. 192–208. doi:10.1080/02626660903546092