Comments by Reviewer 1

[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. We removed the statement related to the symmetrical confidence intervals from the manuscript since we have revised the introduction. We now focus on the differences between Bayesian statistics and frequentist statistics and how both methods compute the confidence intervals of the parameters of the GEV distribution. See page 2 lines 4-11. The motivation of the study is now stated on page 2 lines 21-19.

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. The statement about the bootstrap method to compute the confidence intervals is removed from the manuscript. Please also see the previous comment.

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. See page 5 lines 14-15.

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 (see page 3 lines 2-4). Therefore, the historic flood events have no predictive value without normalizing 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 high detail in Bomers et al. (2019).

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. Why normalization is of high importance is described on page 2 lines 34-35 and page 3 lines 1-2.

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. Please see page 3 lines 2-4.

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 has been added to the revised manuscript. See table 1 on page 4.

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

The following has been 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." See page 4 lines 5-6.

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 an better understandable for decision makers. This argument has been added in the introduction on page 2 lines 25-28. 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 has been 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}." Please see page 11 lines 4-9.

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 76%.

We have added a section to the discussion of the revised manuscript in which it is stated that also for other distribution functions a reduction of the confidence interval was found (page 18 lines 1-7. 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

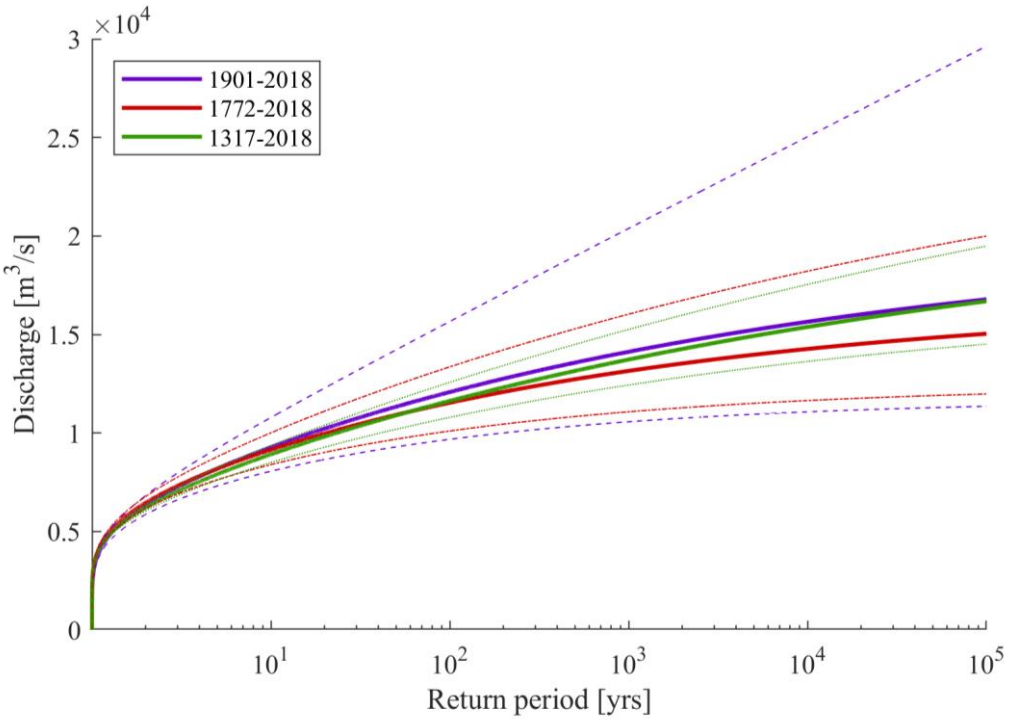
Netherlands at Lobith is limited to a physical maximum value. The crest levels of the dikes along the Lower Rhine are not infinitely high. The height of the dikes influences the discharge capacity of the Lower Rhine and hence the discharge that can flow towards Lobith. This explanation has been added to the revised manuscript (page 11 lines 15-18 and page 12 lines 1-2) such that it becomes clear why we use an upper bounded distribution. The effect of wave overtopping and dike breaches on the discharges at Lobith are explained in high detail by Bomers et al. (2019).

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

Figure 5 has been 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, the entire line has been plotted 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 '$Q_{Bootstrap}$'. This data set has a length equal to the 1317-2018 period. If we compare the line with the 1317-2018 data set, we indeed see that the uncertainty interval of the $Q_{Bootstrap}$ 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 $Q_{Bootstrap}$ 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 1317-2018 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 $Q_{Bootrstrap}$ data set. This explanation has been added to the revised manuscript. Please see page 17 lines 3-10.

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 has been added on page 17 lines 13-14.

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 1901), which correspond with the 1926 flood event. We now refer to Figure 1 to make this clear, see page 14 line 20.

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 m^3/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. This has been added on page 14 line 26-27.

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

We now discuss the most important drawbacks and assumptions of the proposed method. Please see the discussion section. We focus on:

- The added value of normalized historic flood events.
- Resampling the systematic data set
- The use of a single distribution function and goodness-of-fit test
- Length of the extended data set and chosen perception threshold
- Comparison with 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. *Natural Hazards*. DOI: 10.1007/s11069-019-03643-y

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

Comments by reviewer 2

[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."

Please see section 2.2 on page 4 and 5.

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 2 lines 34-35: “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 has been added in the revised manuscript in section 2.3 with in green the new text (see page 5 lines 11-18 and page 6 lines 1-6):

“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 has been added in the revised manuscript (page 6 lines 10-13) 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). 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 is now explained in more detail in the revised manuscript. Please see section 2.3.2 and Figure 3.

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. We have added this in the discussion section. See page 17 lines 21-24.

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. We have now explained step 5 in more detail on page 10 lines 25-27

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. Table 1 has been added to the manuscript 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 has been changed accordingly in the revised manuscript (see table 2).

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 (Figure 1). 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 on page 4 lines 9-12 and page 5 lines 1-2. He found an uncertainty interval of approximately 12%. This has been added explicitly to the revised manuscript to avoid further misunderstanding. Furthermore, we have added a table describing the uncertainties of the various data sets used to extend the systematic data set (see table 1).

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

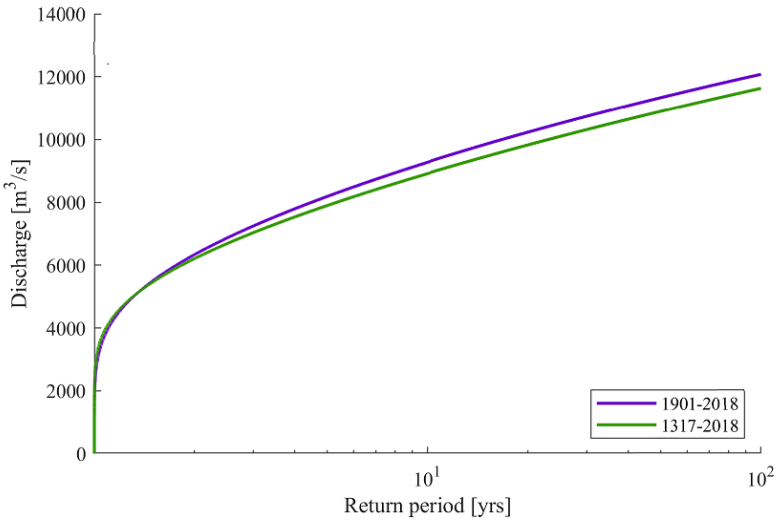
REIS D. S., JR.; STEDINGER J. R. (2005): Bayesian MCMC flood frequency analysis with historical information. In: Journal of Hydrology, 313, pp. 97–116 (cited by the authors)

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). An overview of the uncertainties considered is now given in table 1. The 95% confidence intervals of the flood frequency relations are hence 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. We have added this information in the discussion section on page 19 lines 4-12.



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 have provided a detail explanation of the computed 95% confidence intervals of the following data sets in the revised manuscript (see section 2):

- 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 have added more information about why we propose this method instead of a Bayesian method in the introduction on page 2 lines 4-23. 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 (page 2 lines 24-30).

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 have added a this to the discussion in section 6.5 and the introduction on page 2 lines 14-20.

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 have added the reasons why we set up a bootstrap method in the introduction of the revised manuscript and compared the methodology with the Bayesian statistics briefly in the discussion in section 6.5.

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. 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

Comments by reviewer 3

[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]

In this paper, the authors present a method/case study to reconstruct a continuous times series of annual maximum discharges in order to estimate return times for flood discharges for the Rhine at Lobith. The study uses modern data from 1901 onwards, discharges reconstructed from water level measurements back to 1772 and information from historical flood events back to the 1300. Extending a time series with this information leads to a reduction of uncertainty and to more stable return times. The paper is well structured and written, and the topic is of relevance for flood risk estimation.

However, there general problem I have with this manuscript is that the authors refer to and use data from many other studies, especially the one from Toonen (2015). It is difficult to follow the article for reader if one is not familiar with these studies because it requires reading many secondary sources to gain insight on how all the different data(-sets) were collected and obtained, e.g. how was the regression analysis by Toonen (2015) performed, how were the historical floods in Cologne by Herget and Meurs (2010) reconstructed, etc. This paper includes a lot of different data sets (systematic, historical, plus various bootstrapped time series), it would be beneficial for readers to include a table with a short description and overview of the properties of these data sets and to name them consistently throughout the paper.

Thank you for this remark, we fully agree with you. In the revised manuscript we have provided more knowledge about how the discharges at Lobith were reconstructed by Toonen (2015) (page 4 lines 9-12 and page 5 lines 1-2) as well as the reconstructions at Cologne performed by Herget and Meurs (2010) (page 6 lines 1-6). Furthermore, the following table has been added to the revised manuscript as also suggested by Elena Volpi (first reviewer):

Table 1. Uncertainties and properties of the various data sets used. The 1342-1772 data set represents the historical discharges, whereas the data sets in the period 1772-2018 are referred to as the systematic data set

Time period	Data source	Property	Cause uncertainty	Location
1342-1772	Meurs (2006)	12 single events	Reconstruction uncertain caused by main channel bathymetry, bed friction and maximum occurred water levels	Cologne
1772-1865	Toonen (2015)	Continuous data set	Reconstruction uncertainty based on measured water levels of surrounding sites	Emmerich, Pannerden and Nijmegen
1866-1900	Toonen (2015)	Continuous data set	Uncertainty caused by translation measured water levels into discharges	Lobith
1901-1950	Tijssen (2009)	Continuous data set	Uncertainty caused by extrapolation techniques to translate measured velocities at the water surface into discharges	Lobith
1951-2000	Tijssen (2009)	Continuous data set	Uncertainty caused by translation velocity-depth profiles into discharges	Lobith
2001-2008	Tijssen (2009)	Continuous data set	Measurement errors	Lobith
2009-2018	Measured water levels available at https://waterinfo.rws.nl	Continuous data set	Measurement errors	Lobith

The term “normalize” is used in different contexts (e.g. for historical floods, for the 1900-2008 data set, for the data set of Toonen (2015) which is not normalized but used as normalized data). I find this confusing since it does not become clear what is actually meant by this and what has been done to “normalize” each of these data sets. A more thorough explanation on this matter would be useful.

With the term ‘normalize’ we mean that we translate the historic flood events (water levels, discharges) to present-day discharges at Lobith as a result of changes in the river system and hinterland. Please see also page 2 lines 34-35 where an explanation of the term is given. In the revised manuscript we will explain in more detail how the normalization was done for the various data sets used in this manuscript. The following text has been added with in green te new text:

Regarding the 1901-2008 data set (page 3 lines 16-24):

“Daily discharge observations at Lobith have been performed since 1901 and are available at https://waterinfo.rws.nl. From this data set, the annual maximum discharges are selected in which the hydrologic time period, starting at the 1st of October and ending at the 30th of September, is used. Since changes to the river system have been made the last century, Tijssen (2009) has normalized the measured data set from 1901-2008 to the conditions of the year 2004. In the 20th century, canalization projects were executed along the Upper Rhine (Germany) which were finalized in 1977 (RIZA, 2003). After that, retention measures were executed in the trajectory Andernach-Lobith. Firstly, the 1901-1977 data set has been normalized with the use of a regression function describing the influence of the canalization projects on the maximum discharges. Then, again a regression function was used to normalize the 1901-2008 data set for the retention measures (RIZA, 2003). This results in a normalized 1901-2008 data set for the year 2004.”

Regarding the Toonen (2015) data set (page 5 lines 3-5):

"The reconstructed discharges in the period 1772-1900 represent the computed maximum discharges at the time of occurrence and have not been normalized for changes in the river system and thus they represent the actual occurred annual maximum discharges."

Regarding the Herget and Meurs (2010) data set (page 6 lines 17-19):

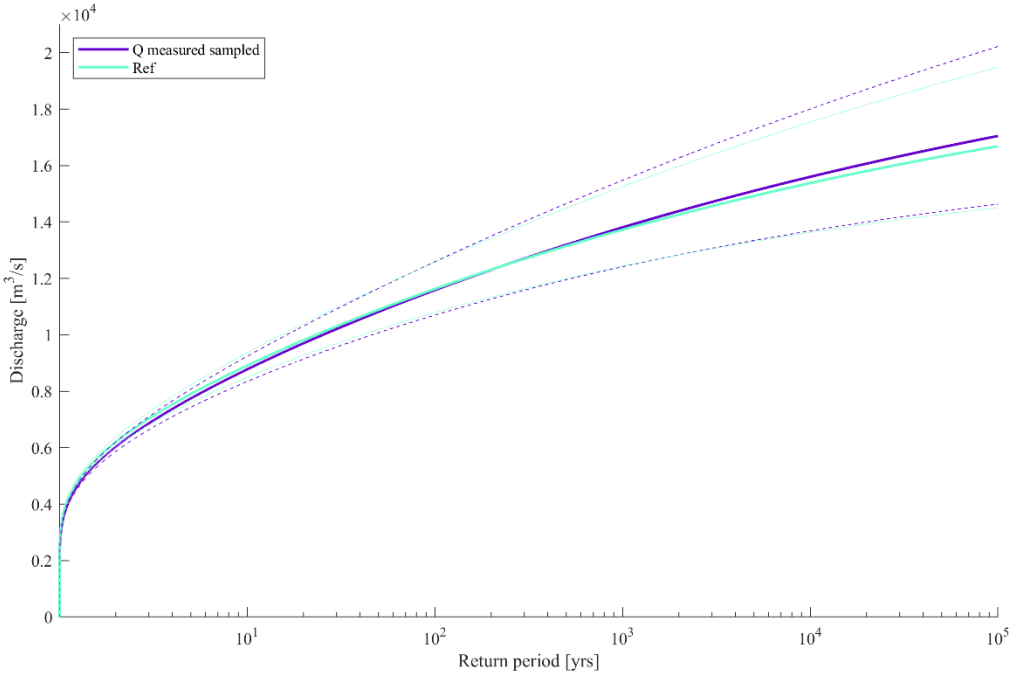
"In this study, the 1D-2D coupled modelling approach as described by Bomers et al. (2019) is used to normalize the data set of Meurs (2006). This normalization is performed by routing the reconstructed historical discharges at Cologne over modern topography to estimate the maximum discharges at Lobith in present times."

In section 2.2 the authors describe the Toonen (2015) data set which uses a linear regression to compute water levels at Lobith. This method leads to a reduced variance of this data set (c.f. table 1). How would this affect the bootstrapping later on, if samples from the so called "systematic time period" with different variances (1772-1900, 1901- 2018) are drawn?

The Toonen (2015) data set indeed has a lower variance compared to the 1901-2018 data set. To identify the effect of using both data sets for resampling purposes, we have performed an additional FFA in which now only the 1901-2018 data set is used for resampling. The results are presented in the figure below in which the purple line indicates the situation in which only the 1901-2018 data set is used for resampling and the blue line represents the reference situation in which the 1772-2018 is used for resampling.

We can see that using the 1772-2018 results in a reduction of the confidence intervals caused by the lower variance in the 1772-1900 data set. This reduction is at maximum 12% for the return period of 100,000 years. This finding has been added to the discussion on page 17 lines 27-34.

However, do note that the lower variance in the 1772-1900 period compared to the 1901-2018 period is most probably a result of natural variability in climate. It is this variability that we want to include in the analysis since also climate variability will exist in the future. If the lower variance was caused by e.g. the removal of a dam construction upstream, it would be reasonable to solely use the 1901-2018 data set for resampling purposes.



From my point of view, the section 2.3.2 presenting the normalization of historical flood events leaves some open questions which need to be addressed. Using a coupled 1D/2D model to route the discharges from Cologne to Lobith seems a reasonable approach given the circumstances of the data, but the dike breach model and the underlying assumptions need more explanation. Is it valid to assume dike breach parameters from today's river geometry for historical times? Is there any historical evidence that there were dike breaches in the past, especially the 1374 event? Especially the reduction of the 1374 flood peak from Cologne to Lobith needs some sound justification/explanation. Why is this reduction only occurring for this specific event? Were there also dike breeches for the other historical events?

Please note that de 1D-2D coupled model is only based on the current geometry and current dike strengths. This is because only then normalization can be performed. So, whether dike breaches occurred during the historical flood events between Andernach and Lobith may be interesting from historical point of view (e.g. a reconstruction of this flood in historical times), but is not directly relevant for this study, as we are interested what will happen nowadays. Therefore, we use so-called fragility curves showing at which water level the dikes in the studied area will start to breach. We now provide more insights in the 1D-2D coupled modelling approach in section 2.3.2 and particularly about the dike breach parameters (please also see figure 3).

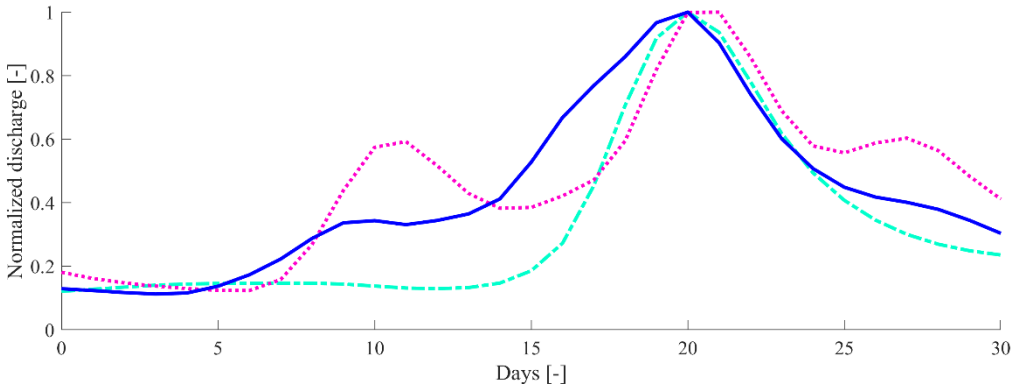
Concerning the 1374 flood event, this event results in a large reduction of the maximum discharge because major overflow and dike breaches occur in present times. Since the 1374 flood event was much larger than the current discharge capacity of the Lower Rhine, the maximum discharge at Lobith decreases. On the other hand, the remaining flood events were below this discharge and hence only a slight reduction in discharges were found for some of the events as a result of dike breaches whereas overflow did not occur. Other events slightly increased as a result of the inflow of the tributaries Sieg, Ruhr and Lippe rivers along the Lower Rhine. This explains why the 1374 flood event is much lower at Lobith compared to the discharge at Andernach, while the discharges of the

remaining flood events are more or less the same at these two locations. This information has been added to the revised manuscript on page 8 lines 21-22 and page 9 lines 1-7.

What exactly is meant by “the upstream discharge shape is varied” (p.6, line 12)? There is a lot of uncertainty in this, which somehow contradicts the aim of the paper to reduce uncertainty.

Of the historic flood events at Cologne, only the peak value was known. The corresponding shape of the discharge wave was unknown. However, this shape may affect the maximum discharge at Lobith. Therefore, we want to include this uncertainty in the analysis. Although it is indeed true that we wanted to reduce uncertainty in flood frequency relations, it does not mean that we want to ignore known uncertainties in the reconstructions.

We used a data set of 250 potential discharge shapes that can occur under current climate conditions (Hegnauer et al., 2014). See the figure below for an example of three potential discharge shapes: e.g. a broad peak, a small peak or a discharge wave with two peaks. For each run in the Monte Carlo analysis, we randomly sampled a shape and scaled this shape to the maximum value of the flood event. This represents the upstream boundary condition of the model run. We now provide more information about the upstream discharge wave shapes on page 7 lines 8-14. Please also see figure 3.



Furthermore, it would be interesting to know if any of 12 historical flood events where winter events, where ice draft/ice jams could/did play a role.

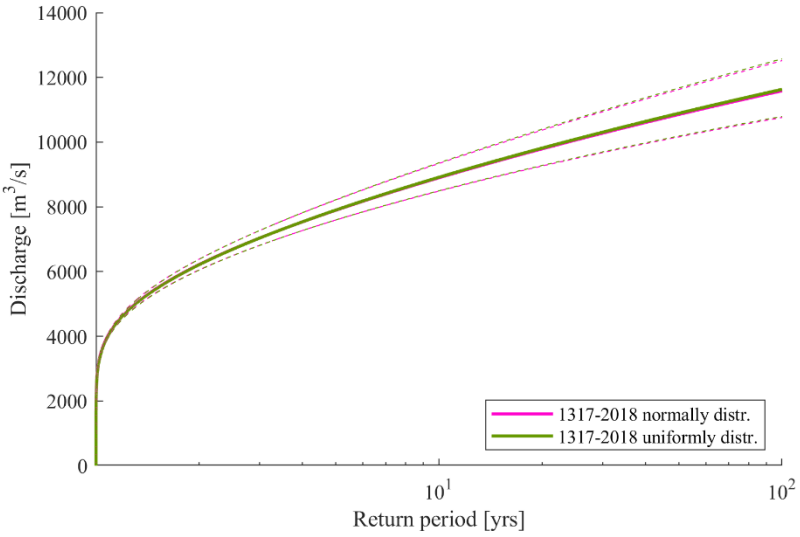
All flood events were winter events, except for the flood event in 1342 that took place in July. However, the flood events caused by ice jams were excluded from the analysis by Herget and Meurs (2010) because of the different hydraulic conditions. All flood events considered are thus caused by high rainfall intensities. This has been added on page 5 lines 12-13.

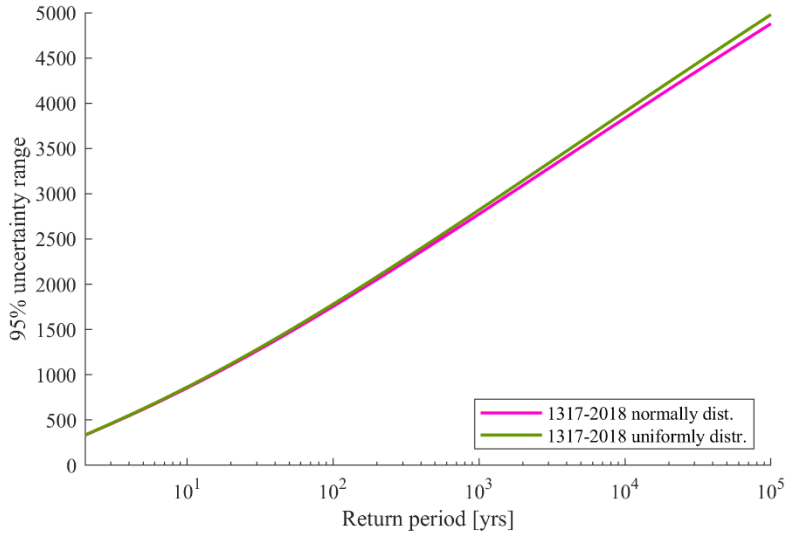
Furthermore, assuming a normal distribution of uncertainties is valid for discharge measurements, but is this also the case for the estimation of historical extreme floods? Or is any discharge values in the uncertainty range equally possible? The reconstruction of the events in Cologne is based on the Manning equation and the uncertainty range results from different roughness coefficients. But do all of these follow a normal distribution?

Herget and Meurs (2010) only provided the maximum, minimum and mean value of the roughness coefficients. They did not provide any insights in the distribution of this uncertainty. We assumed

that they were normal distributed since it is likely that the mean value has a higher probability of occurrence than the boundaries of the considered range. This assumption results in a normal distribution of the maximum discharge at Andernach and consequently to a normal distribution of the maximum discharge at Lobith.

However, we performed the resampling bootstrap method in a different way. During the resampling we assumed uniformly distributed uncertainties and we re-performed the analysis with normally distributed ones. The difference between the two is given in the figure below. We find that assuming normally distributed uncertainties results in slightly smaller uncertainty bounds which can be explained by the lower variance. However, this effect is only very little justifying the assumption of normally distributed uncertainties. This has been added on page 7 lines 3-7.





Section 3: The bootstrap method to create continuous times series is a reasonable approach, however it would also be possible, to use the maximum likelihood method and incorporate the uncertainty range of the historical discharges as well as the discharges lower the perception threshold in the parameter estimation. From my point of view this approach is straight forward and

should yield similar results. Could the authors explain/discuss the benefit of the bootstrapping approach?

We have created a continuous data set by incorporating the uncertainty range of the historical discharges as well as the discharges lower than the perception threshold. Next, we have used the maximum likelihood method to fit each continuous data set (we have 5,000 in total) to a GEV distribution (please see figure 5 and the explanation of the bootstrap method in section 3). We do not understand the difference between our method and the method suggested by the reviewer. If our method was not fully understood by reading the manuscript, we will make this clearer in the revised manuscript.

In Section 4, the authors state that there are many distributions and fitting methods for flood frequency analysis and that the only use the GEV with maximum likelihood method. It seems justified, that only one combination is used to quantify the reduction of the uncertainty, but in practice there are many different distributions and parameter estimation methods - which again cause higher uncertainties in the estimation of return times, especially for the upper tail extremes. The authors should include a comment and if possible a quantification of this effect on this in the discussion.

You are indeed correct that the use of various kinds of distributions and parameter estimation methods influence the uncertainty in the flood frequency relations. We have performed the analysis with the Gumbel and Weibull distribution as well and these results are now shown in the discussion. We will also highlight that using the combination of multiple distributions in the analysis increases the uncertainties in the estimation of return periods. We have added this in the discussion of the revised manuscript in section 6.3.

In Section 5.2., the authors argue that the reconstruction of historical flood events is complicated and time consuming and that this can be overcome by bootstrapping. However, the information from rare and large historical flood events is still required as is stated at the end of the section. This sounds like an inconsistency in the line of argumentation. Furthermore, this whole section is somewhere between results and discussion. I suggest that the authors try to separate more clearly between results and discussion.

We indeed argue that reconstructing historic flood events is time consuming. Therefore, we studied whether it is also possible to only use the 1901-2018 measured data set in a bootstrap approach. However, we find that the uncertainty interval of this FF curve is larger than for the FF curve in which the normalized historic flood events are considered. We thus show that, although it is time consuming to normalize the historic flood events, it is worth the effort since it reduces uncertainties in FF relations. Since this was not fully clear, we have rewritten the paragraph in the revised manuscript. Furthermore, we have rewritten the paragraph in a more discussion style and replace this section towards the discussion section. Please see section 6.1.

In the discussion, the effect of a hypothetical future extreme flood on the robustness of return times is addressed, which is somehow obvious from my point of view. This aspect does not add much value to the paper and can either be omitted or be moved to the results section.

We have moved this section towards the results. We believe that it shows the robustness of the method since using an extended data set in flood management avoids that a flood frequency curve changes after the occurrence of a future flood event. As a results, the FFA does not have to be performed again, while this is necessary if only the data set of measured discharges is used. Therefore, decision makers might have taken another decision. Please see page 14 lines 26-27.

Some specific comments:

Page 3, line 3f.: Why are uncertainties not symmetrical due to missing continuous data? Don't these result from the non-linearity of the rating curves?

The sentence about the symmetrical uncertainties stated in the introduction was not fully correct. Indeed, uncertainties are in general not symmetrical for flood frequency relations. This is indeed the result of the non-linearity of the rating curves. However, the introduction have been revised significantly and as a result the sentence related to the symmetrical uncertainties have been removed.

Page 4, line 7f.: ACDP-measurements are in general not free of uncertainties, this assumption is not correct.

Indeed, the ACDP-measurements are in general not free of uncertainties. Since we had no reference regarding this uncertainty, we used the uncertainties as suggested by Toonen (2015). He mentioned that only the discharges slightly exceeding the bank-full discharge have an uncertainty range of 5%. In the revised manuscript we now include this uncertainty for all ACDP-measurements (see page 3 lines 31-32 and table 1). However, since all annual maximum discharges in the period 2000-2018 where between 4,000 and 8,000 m^3/s, the 5% uncertainty was already included in the analysis and hence the results will not change.

Page 11, line 2: Where does this confidence interval of 7400m3 /s come from?

This value represents the reduction in the confidence interval if the 1901 data set is extended towards 1317 for the discharges corresponding with a return period of 1,250 years. We have rewritten the text such that this becomes clearer. Please see page 12 line 22 and page 13 lines 1-2.

Page 15, line 1: Same as above, modern discharge measurements are not free of measurement errors!

Please see above and page 3 lines 31-32 and table 1.

Page 15, line 5f.: See above, this is not a novel results and can more or less be expected. Furthermore, the statement that "flood managers can be less nervous" sounds awkward and is not really correct, since the uncertainty caused by different distributions/parameter estimation methods is not addressed.

This section has been moved to the results. It is indeed true that we did not include the uncertainty caused by different distributions and parameter estimation methods. We have removed the statement from the manuscript and added in the discussion the effects of using a combination of different distributions on the uncertainty intervals. Please see section 6.3.

Figure 2: Should be replaced by a "conventional" map, including national boundaries, a scale bar etc. Readers from outside of Europe might not be familiar with this region.

The figure has been replaced by the following figure such that now the national boundaries, scale bar, north arrow, names and model boundary are given. Please see figure 2

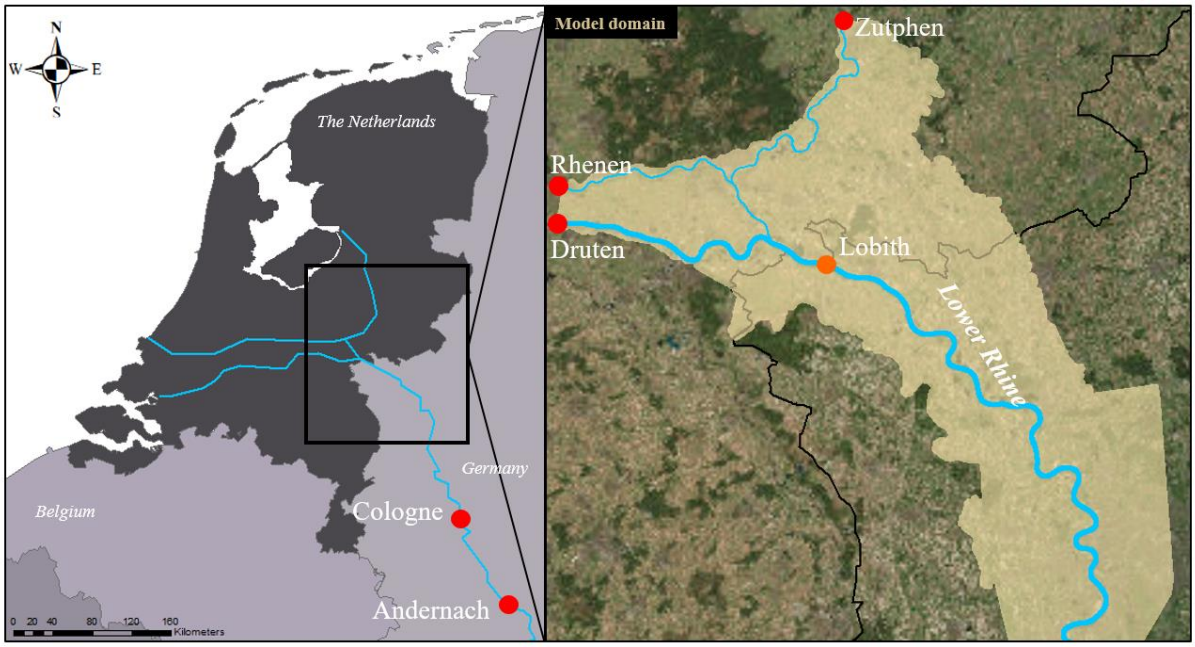
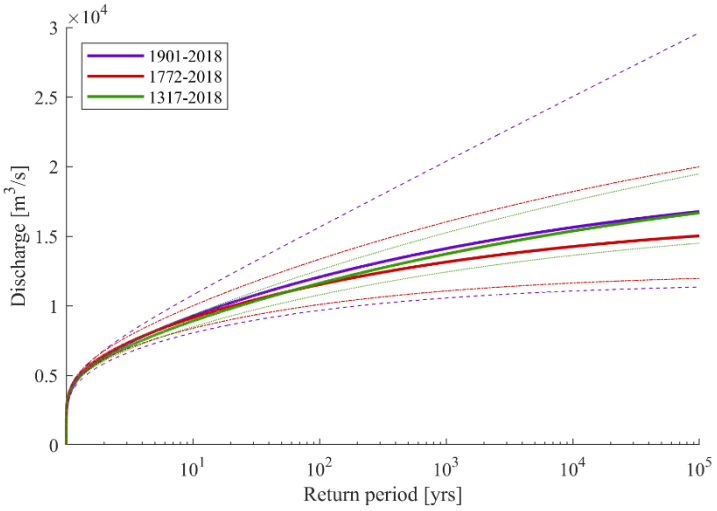


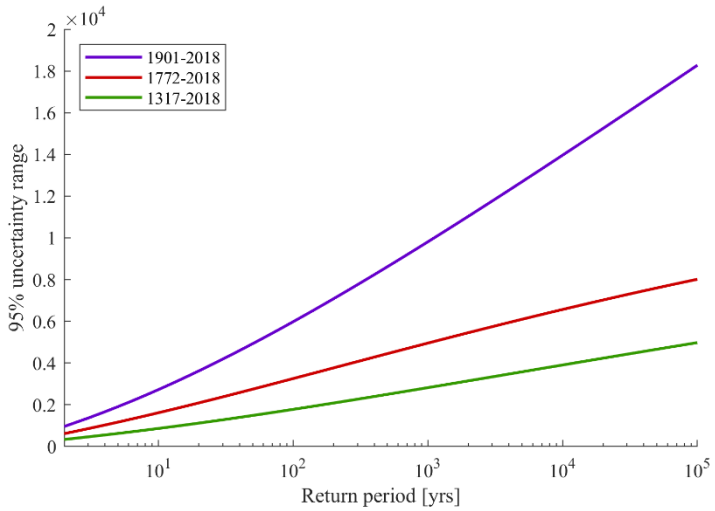
Table 1: The results of Toonen (2015) can be omitted in this table from my point of view.

The results of Toonen (2015) are omitted from the table.

Figure 6 and 7: The colours/line styles of the different curves are difficult to distinguish and should be changed to make these figures better to read.

The colours are adapted as follow:





References: To my knowledge, Meurs 2006 is a diploma thesis, not a PhD thesis.

You are correct, it is indeed a diploma thesis. This has been adapted in the revised manuscript.

REFERENCES:

Hegnauer, M., Beersma, J. J., van den Boogaard, H. F. P., Buishand, T. A., and Passchier, R. (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

List of all relevant changes made in de manuscript

- Introduction has significantly been changed. We now provide more information about the differences between the Bayesian approach and the frequentist statistics. Also the reasons why we set up a bootstrap method are now better described.
- More information about the data sets used is given as well as the uncertainties involved.
- More information about the hydraulic model has been added in which we now describe in more detail how the normalization steps of the historical flood events are performed.
- The discussion section has been altered a lot. More information about the assumptions and drawbacks are given.

Decreasing uncertainty in flood frequency analyses by including historic flood events in an efficient bootstrap approach

Anouk Bomers[1], Ralph M. J. Schielen[1,2], and Suzanne J. M. H. Hulscher[1]

[1]University of Twente, Dienstweg 1, Enschede, The Netherlands
[2]Ministry of Infrastructure and Water Management-Rijkswaterstaat, Arnhem, The Netherlands

Correspondence: A. Bomers (a.bomers@utwente.nl)

Abstract. Flood frequency curves are usually highly uncertain since they are based on short data sets of measured discharges or weather conditions. To decrease the confidence intervals, an efficient bootstrap method is developed in this study. The Rhine river delta is considered as a case study. We use a hydraulic model to normalize historic flood events for anthropogenic and natural changes in the river system. As a result, the data set of measured discharges could be extended with approximately 600 years. The study shows that flood events decrease the confidence interval of the flood frequency curve significantly, specifically in the range of large floods. This even applies if the maximum discharges of these historic flood events are highly uncertain themselves.

1 Introduction

Floods are one of the main natural hazards to cause large economic damage and human casualties worldwide as a result of serious inundations with disastrous effects. Design discharges associated with a specific return period are used to construct flood defences to protect the hinterland from severe floods. These design discharges are commonly determined with the use of a flood frequency analysis (FFA). The basic principle of an FFA starts with selecting the annual maximum discharges of the measured data set, or peak values that exceed a certain threshold (Schendel and Thongwichian, 2017). These maximum or peak values are then used to identify the parameters of a probability distribution. From this fitted distribution, discharges corresponding to any return period can be derived.

Return periods of design discharges are commonly in the order of 500 years or even more, while discharge measurements have been performed only for the last 50-100 years. For the Dutch Rhine river delta (used as a case study in this paper), water levels and related discharges have been registered since 1901 while design discharges have a return period up to 100,000 years (Van der Most et al., 2014). Extrapolation of these measured discharges to such return periods results in large confidence intervals of the predicted design discharges. Uncertainty in the design discharges used for flood risk assessment can have major implications for national flood protection programs since it determines whether and where dike reinforcements are required. A too wide uncertainty range may lead to unnecessary investments.

To obtain an estimation of a flood with a return period of e.g. 10,000 years with little uncertainty, a discharge data set of at least 100,000 years is required (Klemeš, 1986). Of course, such data sets do not exist. For this reason, many studies try to extend

the data set of measured discharges with historic and/or paleo flood events. The most common methods in literature to include historical data into an FFA are based on the traditional methods of frequentist statistics (Frances et al., 1994; MacDonald et al., 2014; Sartor et al., 2010) and Bayesian statistics (O'Connell et al., 2002; Parkes and Demeritt, 2016; Reis and Stedinger, 2005).

While frequentist statistics are generally applied by decision makers, Bayesian statistics have significantly increased in popularity in the last decade. Reis and Stedinger (2005) has successfully applied a Bayesian Markov Chain Monte Carlo (MCMC) analysis to determine flood frequency relations and their uncertainties using both systematic data and historic flood events. A Bayesian analysis determines the full posterior distribution of the parameters of a probability distribution function (e.g. GEV distribution). This has as advantage that the entire range of parameter uncertainty can be included in the analysis. Contrarily, classical methods based on frequentist statistics usually only provide a point estimate of the parameters where after their uncertainties are commonly described by using the assumption of symmetric normal distributed uncertainty intervals (Reis and Stedinger, 2005). The study of Reis and Stedinger (2005) shows that confidence intervals of design discharges were reduced significantly by extending the systematic data set with historic events using the proposed Bayesian framework. This finding is important for the design of future flood reducing measures since these can then be designed with less uncertainty.

However, Bayesian statistic also has several drawbacks. Although no assumption about the parameter uncertainty of the distribution function has to be made, the results depend on the parameter priors which have to be chosen a priori. The influence of the priors on the posterior distributions of the parameters and hence on the uncertainty of flood frequency relations can even be larger than the influence of discharge measurement errors (Neppel et al., 2010). The prior can be estimated by fitting the original data with the use of e.g. the Maximum Likelihood method. However, we do not have any measurements in, or near to, the tail of the frequency distribution functions. In this way, the benefits of the Bayesian method compared to a traditional flood frequency analysis are at least questionable.

In this study, we propose a systematic approach to include historic flood information into flood safety assessments. The general methodology of a flood frequency analysis remains, only the data set of measured discharges is extended with the use of a bootstrap approach. As a result, this method is close to current practice of water managers. We extend the data set of measured discharges at Lobith, the German-Dutch border, with historic events to decrease uncertainty intervals of design discharges corresponding to rare events. A bootstrap method is proposed to create a continuous data set after which we perform a traditional FFA to stay in line with the current methods used for Dutch water policy. Hence, the results are well understandable by decision makers since solely the effect of using data sets with different lengths on flood frequency relations and corresponding uncertainty intervals are presented. The objective of this study is thus to develop a straightforward method to consider historic flood events in an FFA, while the basic principles of an FFA remain unchanged.

The measured discharges at Lobith (1901-2018) are extended with the continuous reconstructed data set of Toonen (2015) covering the period 1772-1900. These data sets are extended with the most extreme, older historic flood events near Cologne reconstructed by Meurs (2006), which are routed towards Lobith. For this routing, a one dimensional-two dimensional (1D-2D) coupled hydraulic model is used to determine the maximum discharges during these historic events based on the current geometry. 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. Normalizing the historic events is of high importance since flood patterns

most likely change over the years as a result of e.g. dike reinforcements, land use change or decrease in floodplain area (dike shifts). The normalized events almost always lead to a higher discharge than the historic event. 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. In any case, the normalized events give insight in the consequences of an event with the same characteristics of a historic flood event translated to present times. To create a continuous data set, a bootstrap resampling technique is used. The results of the bootstrap method are evaluated against an FFA based on solely measured annual maximum discharges (1901-2018 and 1772-2018). Specifically, the change in the design discharge and its 95% confidence interval of events with a return period of 100,000 years is considered because this design discharge corresponds with the highest safety level used in Dutch flood protection programs (Van Alphen, 2016).

In Section 2 the different data sets used to construct the continuous discharge data set are explained, as well as the 1D-2D coupled hydraulic model. Next, the bootstrap method and FFA are explained (Section 3 and Section 4 respectively). After that, the results of the FFA are given (Section 5). The paper ends with a discussion (Section 6) and the main conclusions (Section 7).

2 Annual maximum discharges

2.1 Discharge measurements period 1901 - present

Daily discharge observations at Lobith have been performed since 1901 and are available at https://waterinfo.rws.nl. From this data set, the annual maximum discharges are selected in which the hydrologic time period, starting at the 1st of October and ending at the 30th of September, is used. Since changes to the system have been made the last century, Tijssen (2009) has normalized the measured data set from 1901-2008 for the year 2004. In the 20th century, canalization projects were executed along the Upper Rhine (Germany) which were finalized in 1977 (Van Hal, 2003). After that, retention measures were executed in the trajectory Andernach-Lobith. Firstly, the 1901-1977 data set has been normalized with the use of a regression function describing the influence of the canalization projects on the maximum discharges. Then, again a regression function was used to normalize the 1901-2008 data set for the retention measures (Van Hal, 2003). This results in a normalized 1901-2008 data set for the year 2004. For the period 2009-2018, the measured discharges without normalization are used.

During the discharge recording period, different methods have been used to perform the measurements. These different methods result in different uncertainties (Table 1 and must be included in the FFA to correctly predict the 95% confidence interval of the FF curve. From 1901 until 1950, discharges at Lobith were based on velocity measurements performed with floating sticks on the water surface. Since the velocity was only measured at the surface, extrapolation techniques were used to compute the total discharge. This resulted in an uncertainty of approximately 10% (Toonen, 2015). From 1950 until 2000, current meters were used to construct velocity-depth profiles. These profiles were used to compute the total discharge, having an uncertainty of approximately 5% (Toonen, 2015). Since 2000, Acoustic Doppler Current Profiles have been used for which also an uncertainty of 5% is assumed.

Table 1. Uncertainties and properties of the various data sets used. The 1342-1772 data set represents the historical discharges (first row in the table), whereas the data sets in the period 1772-2018 are referred to as the systematic data set (rows 2-7)

Time period	Data source	Property	Cause uncertainty	Location
1342-1772	Meurs (2006)	12 single events	Reconstruction uncertain caused by main channel bathymetry, bed friction and maximum occurred water levels	Cologne
1772-1865	Toonen (2015)	Continuous data set	Reconstruction uncertainty based on measured water levels of surrounding sites ($\sim 12\%$)	Emmerich, Pannerden and Nijmegen
1866-1900	Toonen (2015)	Continuous data set	Uncertainty caused by translation measured water levels into discharges ($\sim 12\%$)	Lobith
1901-1950	Tijssen (2009)	Continuous data set	Uncertainty caused by extrapolation techniques to translate measured velocities at the water surface into discharges (10%)	Lobith
1951-2000	Tijssen (2009)	Continuous data set	Uncertainty caused by translation velocity-depth profiles into discharges (5%)	Lobith
2001-2008	Tijssen (2009)	Continuous data set	Measurement errors (5%)	Lobith
2009-2018	Measured water levels available at https://waterinfo.rws.nl	Continuous data set	Measurement errors (5%)	Lobith

2.2 Water level measurements period 1772 - 1900

Toonen (2015) studied the effects of non-stationarity in flooding regimes over time on the outcome of an FFA. He extended the data set of measured discharges of the Rhine river at Lobith with the use of water level measurements. 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, Germany (located 10 kilometers in upstream direction), Pannerden (located 10 kilometers in downstream direction) and Nijmegen (located 22 kilometers in downstream direction). Toonen (2015) used the water levels of these locations to compute the water levels at Lobith and their associated uncertainty interval with the use of a 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 level 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

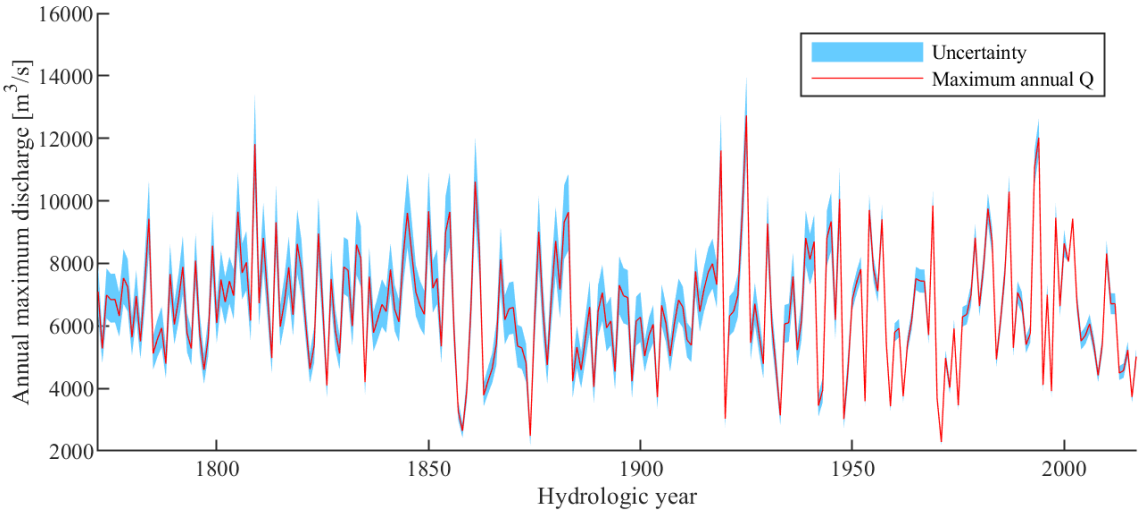


Figure 1. Maximum discharges (Q) and their 95% confidence interval during the systematic time period (1772-2018)

at Lobith. 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) (Toonen, 2015).

The reconstructed discharges in the period 1772-1900 represent the computed maximum discharges at the time of occurrence and these have not been normalized for changes in the river system. They thus represent the actual occurred annual maximum discharges. Toonen (2015) argues that, based on the work of Bronstert et al. (2007) and Vorogushyn and Merz (2013), the effect of recent changes in the river system on discharges of extreme floods of the Lower Rhine is small. Hence, it is justified to use the presented data set of Toonen (2015) in this study as normalized data. Fig. 1 shows the annual maximum discharges for the period 1772-2018 and their 95% confidence intervals. This data represents the systematic data set and consists of the measured discharges covering the period 1901-2018 and the reconstructed data set of Toonen (2015) covering the period 1772-1900.

2.3 Reconstructed flood events period 1300 AD - 1772

Meurs (2006) has reconstructed maximum discharges during historic flood events near the city of Cologne, Germany. The oldest event dates back to 1342. Only flood events caused by high rainfall intensities were reconstructed because of the different hydraulic conditions of flood events caused by ice jams. 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 levels 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 Manning's equation:

$$Q_{\mathrm{p}} = A_{\mathrm{p}} {R_{\mathrm{p}}}^{2/3} S^{1/2} n^{-1} \quad (1)$$

where Q_p represents the peak discharge (m^3/s), A_p the cross-sectional area (m^2) during the highest flood level, R_p the hydraulic radius during the highest flood level (m), S the slope of the main channel and n its Manning's roughness coefficient ($s/m^{1/3}$). However, the highest flood level as well as the Manning's roughness coefficient are uncertain. The range of maximum water levels were based on historical sources, whereas the range of Manning's roughness coefficients were based on the tables of Chow (1959). Including these uncertainties in the analysis, Herget and Meurs (2010) were able to calculate maximum discharges of the specific historic flood events and associated uncertainty ranges (Fig. 4).

In total 13 historic flood events that occurred before 1772 were reconstructed. Two of the flood events occurred in 1651. Only the largest flood of these two is considered as data point. This results in 12 historic floods that are used to extend the systematic data set. The reconstructed maximum discharges at Cologne (Meurs, 2006) are used to predict maximum discharges at Lobith with the use of a hydraulic model to normalize the data set. Although Cologne is located roughly 160 km upstream of Lobith, there is a strong correlation between the discharges at these two locations. This is because they are located in the same fluvial trunk valley and only have minor tributaries (Sieg, Ruhr and Lippe rivers) joining in between (Toonen, 2015). This makes the reconstructed discharges at Cologne applicable to predict corresponding discharges at Lobith. The model used to perform the hydraulic calculations is described in Section 2.3.1. The maximum discharges at Lobith of the 12 historic flood events are given in Section 2.3.2.

2.3.1 Model environment

In this study, the 1D-2D coupled modelling approach as described by Bomers et al. (2019a) is used to normalize the data set of Meurs (2006). This normalization is performed by routing the reconstructed historical discharges at Cologne over modern topography to estimate the maximum discharge at Lobith in present times. The study area stretches from Andernach to the Dutch cities of Zutphen, Rhenen and Druten (Fig. 2). In the hydraulic model, the main channels and floodplains are discretized by 1D profiles. The hinterland is discretized by 2D grid cells. The 1D profiles and 2D grid cells are connected by a structure corresponding with the dimensions of the dike that protects the hinterland from flooding. If the computed water level of a 1D profile exceeds the dike crest, water starts to flow into the 2D grid cells corresponding with inundations of the hinterland. A discharge wave is used as upstream boundary condition. Normal depths, computed with the use of the Manning's equation, were used as downstream boundary conditions. HEC-RAS (v. 5.0.3) (Brunner, 2016), developed by the Hydrologic Engineering Centre (HEC) of the US Army Corps of Engineers, is used to perform the computations. For more information about the model set-up, see Bomers et al. (2019b).

2.3.2 Normalization historic flood events

We use the hydraulic model to route the historical discharges at Cologne, as reconstructed by Meurs (2006), to Lobith. However, the reconstructed historical discharges were uncertain. Therefore, also the discharges at Lobith are uncertain. To include this uncertainty in the analysis a Monte Carlo analysis (MCA) is performed in which, among others, the upstream discharges reconstructed by Meurs (2006) are included as random parameters. These discharges have large confidence intervals (Fig. 4). The severe 1374 flood, representing the largest flood of the last 1,000 years with a discharge of 23,000 m^3/s, even has a

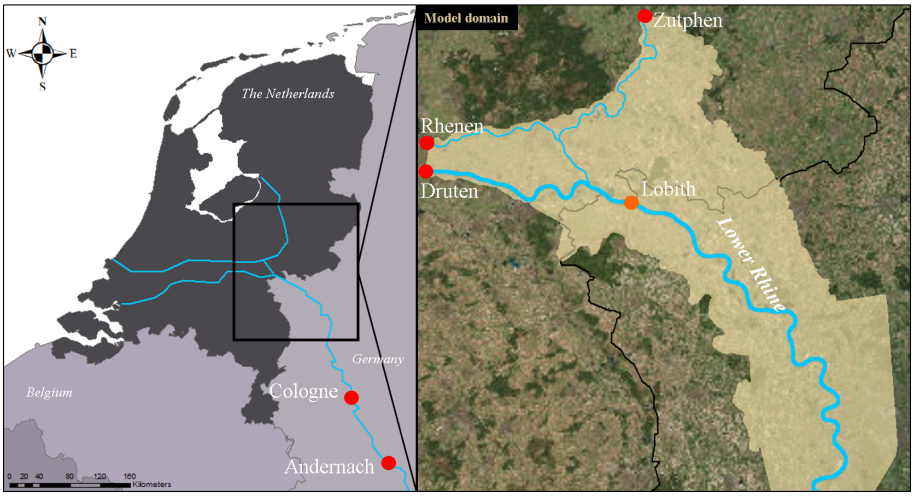


Figure 2. Model domain (blue river branches) of the 1D-2D coupled model

confidence interval of more than 10,000 m^3/s. To include the uncertainty as computed by Meurs (2006) in the analysis, the maximum upstream discharge is varied in the MCA based on its probability distribution. However, the shape of this probability distribution is unknown. Herget and Meurs (2010) only provided the maximum, minimum and mean value of the reconstructed discharges. We assumed normally distributed discharges since it is likely that the mean value has a higher probability of occurrence than the boundaries of the reconstructed discharge range. However, we found that the assumption of the uncertainty distribution has a negligible effect on the 95% uncertainty interval of the FF curve at Lobith. Assuming uniformly distributed uncertainties only led to a very small increase in this 95% uncertainty interval.

Not only the maximum discharges at Cologne are uncertain, also the discharge wave shape of the flood event. The shape of the upstream flood event may influence the maximum discharge at Lobith. Therefore, the upstream discharge wave shape is varied in the MCA. We use a data set of approximately 250 potential discharge wave shapes that can occur under current climate conditions (Hegnauer et al., 2014). In such a way, a broad range of potential discharge wave shapes, e.g. a broad peak, a small peak, or two peaks, are included in the analysis. For each run in the MCA, a discharge wave shape is randomly sampled and scaled to the maximum value of the flood event considered (Fig. 3). This discharge wave represents the upstream boundary condition of the model run.

The sampled upstream discharges, based on the reconstructed historic discharges at Cologne, may lead to dike breaches in present times. Since we are interested in the consequences of the historic flood events in present times, we want to include these dike breaches in the analysis. However, it is highly uncertain how dike breaches develop. Therefore, the following potential dike breach settings are included in the MCA (Fig. 3):

1. Dike breach threshold
2. Final dike breach width

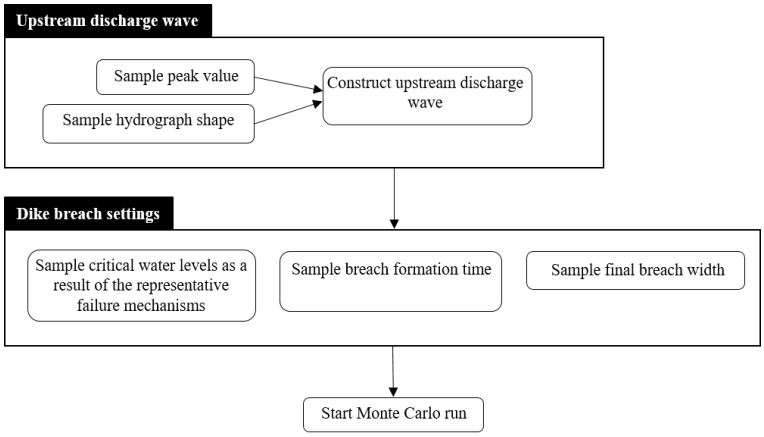


Figure 3. Random input parameters considered in the Monte Carlo analysis

3. Dike breach duration

The dike breach thresholds (i.e. the critical water level at which a dike starts to breach) are based on 1D fragility curves provided by the Dutch Ministry of Infrastructure and Water Management. A 1D fragility curve expresses the reliability of a flood defence as a function of the critical water level (Hall et al., 2003). The critical water levels thus influence the timing of dike breaching. For the Dutch dikes, it is assumed that the dikes can fail due to failure mechanisms wave overtopping and overflow, piping and macro-stability, where as the German dikes only fail because of wave overtopping and overflow (Bomers et al., 2019b). The distributions of the final breach width and the breach formation time are based on literature and on historical data (Apel et al., 2008; Verheij and Van der Knaap, 2003). Since it is unfeasible to implement each dike kilometer as potential dike breach location in the model, only the dike breach locations that result in significant overland flow are implemented. This results in 33 potential dike breach locations whereas overflow (without dike breaching) is possible to occur at every location throughout the model domain (Bomers et al., 2019b).

So, for each Monte Carlo run an upstream maximum discharge and discharge wave shape is sampled. Next, for each of the 33 potential dike breach locations the critical water level, dike breach duration and final breach widths are sampled. With this data, the Monte Carlo run representing a specific flood scenario can be run (Fig. 3). This process is repeated until converge of the maximum discharge at Lobith and its confidence interval is found. For a more in depth explanation of the Monte Carlo analysis and random input parameters, we refer to Bomers et al. (2019b).

The result of the MCA is the normalized maximum discharge at Lobith and its 95% confidence interval for each of the 12 historic flood events. Since the maximum discharges at Cologne are uncertain, also the normalized maximum discharges at Lobith are uncertain (Fig. 4). Fig 4 shows that the extreme 1374 flood with a maximum discharge of between 18,800 m^3/s and 29,000 m^3/s at Cologne, reduces significantly in downstream direction as a result of overflow and dike breaches. Consequently, the maximum discharge at Lobith turns out to be between 13,825 and 17,753 m^3/s. This large reduction in the maximum discharge is caused by the major overflow and dike breaches that occur in present times. Since the 1374 flood event was much

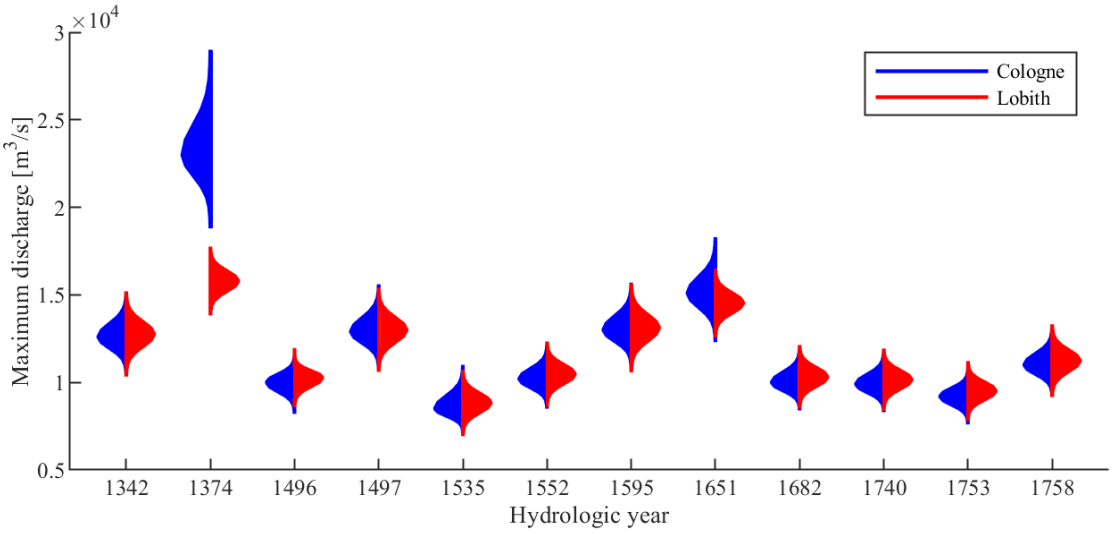


Figure 4. Maximum discharges and their 95% confidence intervals of the reconstructed historic floods at Cologne (Herget and Meurs, 2010) and simulated maximum discharges and their 95% confidence intervals at Lobith for the 12 historic flood events

larger than the current discharge capacity of the Lower Rhine, the maximum discharge at Lobith decreases. The reconstruction of the 1374 flood over modern topography is presented in detail in Bomers et al. (2019c). On the other hand, the other 11 flood events were below this discharge and hence only a slight reduction in discharges was found for some of the events as a result of dike breaches whereas overflow did not occur. Some other events slightly increased as a result of the inflow of the tributaries Sieg, Ruhr and Lippe rivers along the Lower Rhine. This explains why the 1374 flood event is much lower at Lobith compared to the discharge at Andernach, while the discharges of the other 11 flood events are more or less the same at these two locations (Fig. 4). The reduction in maximum discharge of the 1374 flood event in downstream direction shows the necessity to apply hydraulic modelling since the use of a linear regression analysis based on measured discharges between Cologne and Lobith will result in an unrealistic larger maximum discharge at Lobith.

The reconstructed discharges at Lobith are used to extend the systematic data set presented in Fig. 1. In the next section, these discharges are used in an FFA with the use of a bootstrap method.

3 Bootstrap method

The systematic data set covering the period 1772-2019 is extended with 12 reconstructed historic flood events that occurred in the period 1300-1772. To create a continuous data set, a bootstrap method based on sampling with replacement is used. The continuous systematic data set (1772-2018) is resampled over the missing years from the start of the historical period to the start of the systematic record. Two assumptions must be made such that the bootstrap method can be applied:

1. The start of the continuous discharge series since the true length of the historical period is not known.

2. The perception threshold over which floods were recorded in the historical times before water level and discharge measurements were conducted.

Assuming that the historical period starts with the first known flood (in this study: 1342) will significantly underestimate the true length of this period. This underestimation influences the shape of the FF curve (Hirsch and Stedinger, 1987; Schendel and Thongwichian, 2017). Therefore, Schendel and Thongwichian (2017) proposed the following equation to determine the length of the historical period:

$$M = L + \frac{L + N - 1}{k} \tag{2}$$

where M represents the length of the historical period (years), L the number of years from the first historic flood to the start of the systematic record (431 years), N the length of the systematic record (247 years) and k the number of floods exceeding the perception threshold in both the historical period as well as in the systematic record (28 in total). Using equation 2 results in a length of the historical period of 455 years (1317-1771).

The perception threshold is considered to be equal to the discharge of the smallest flood present in the historic period, representing the 1535 flood with an expected discharge of 8,826 m^3/s (Fig. 4). We follow the method of Parkes and Demeritt (2016) assuming that the perception threshold was fairly constant over the historical period. However, the maximum discharge of the 1535 flood is uncertain and hence also the perception threshold is uncertain. Therefore, the perception threshold is treated as a random uniformly distributed parameter in the bootstrap method which boundaries are based on the 95% confidence interval of the 1535 flood event.

The bootstrap method consist of creating a continuous discharge series from 1317-2018. The method includes the following steps (Fig. 5):

1. Combine the 1772-1900 data set with the 1901-2018 data set to create a systematic data set.
2. Select the flood event with the lowest maximum discharge present in the historic time period. Randomly sample a value in between the 95% confidence interval of this lowest flood event. This value is used as perception threshold.
3. Compute the start of the historical time period (equation 2).
4. Of the systematic data set, select all discharges that have an expected value lower than the sampled perception threshold.
5. Use the data set created in Step 4 to create a continuous discharge series in the historical time period. Randomly draw an annual maximum discharge of this systematic data set for each year within the historical period of which no data is available following a bootstrap approach.
6. Since both the reconstructed as well as the measured discharges are uncertain due to e.g. measurement errors, these uncertainties must be included in the analyses. Therefore, for each discharge present in the systematic data set and in the historical data set, its value is randomly sampled based on its 95% confidence interval.
7. Combine the data sets of Steps 5 and 6 to create a continuous data set starting from 1317-2018.

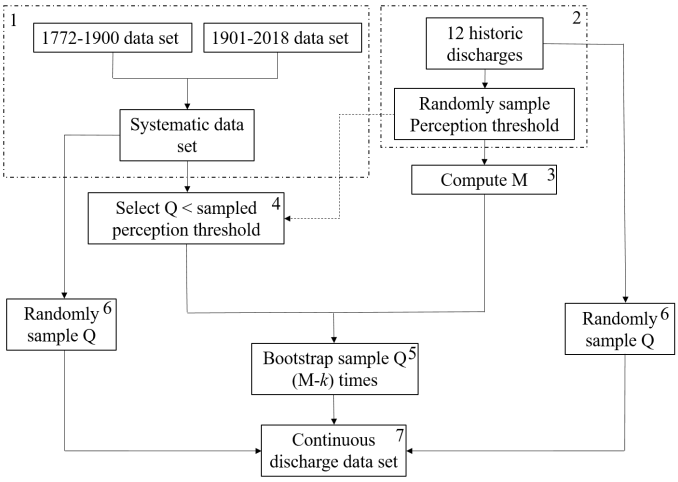


Figure 5. Bootstrap method to create a continuous discharge series

The presented steps in the bootstrap method are repeated 5,000 times in order to create 5,000 continuous discharge data sets resulting in convergence in the FFA. The FFA procedure itself is explained in the next section.

4 Flood frequency analysis

An FFA is performed to determine the FF relation of the different data sets (e.g. systematic record, historical records). 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 since the safety standards expressed in probability of flooding along the Dutch dikes vary from 10^{-2} to 10^{-5}. Furthermore, we highlight that we focus on the influence of extending the data set of measured discharges on the reduction in uncertainty of the FF relations rather than on the suitability of the different distributions and fitting methods.

We restrict our analysis to the use of a Generalized Extreme Value (GEV) distribution since this distribution is commonly used in literature to perform an FFA (Parkes and Demeritt, 2016; Haberlandt and Radtke, 2014; Gaume et al., 2010). Additionally, several studies have shown the applicability of this distribution on the flooding regime of the Rhine river (Toonen, 2015; Chbab et al., 2006; Te Linde et al., 2010). The GEV distribution has an upper bound and is thus capable of flattening off at extreme values by having a flexible tail. We use a bounded distribution since the maximum discharge that is capable of entering the Netherlands is limited to a physical maximum value. The crest levels of the dikes along the Lower Rhine, Germany, are not infinitely high. The height of the dikes influences the discharge capacity of the Lower Rhine and hence the discharge that can

Table 2. Discharges [m^3/s] and their 95% confidence interval corresponding to several return periods for the 1901, 1772 and 1317 data sets and the data set of Toonen (2015)

Data	Q_{10}	Q_{100}	$Q_{1,000}$	2.5%	$Q_{1,250}$	97.5%	2.5%	$Q_{100,000}$	97.5%
1901-2018	9,264	12,036	14,050	10,594	14,215	20,685	11,301	16,649	29,270
1772-2018	9,106	11,442	13,008	11,053	13,130	16,027	11,858	14,813	19,576
1317-2018	8,899	11,585	13,655	12,514	13830	15,391	14,424	16,562	19,303

flow towards Lobith. Using an upper bounded distribution yields that the FF relation converges towards a maximum value for extremely large return periods. This value represents the maximum discharge that is capable of occurring at Lobith.

The GEV distribution is described with the following equation:

$$F(x) = exp\{-[\xi\frac{x-\mu}{\sigma}]^{\frac{1}{\xi}}\} \tag{3}$$

where μ represents the location parameter indicating where the origin of the distribution is positioned, σ the scaling parameter describing the spread of the data, and ξ represents the shape parameter controlling the skewness and kurtosis of the distribution, both influencing the upper tail and hence the upper bound of the system. The maximum likelihood method is used to determine the values of the three parameters of the GEV distribution (Stendinger and Cohn, 1987; Reis and Stedinger, 2005).

The FFA is performed for each of the 5,000 continuous discharge data sets created with the bootstrap method (Section 3), resulting in 5,000 fitted GEV curves. The average of these relations is taken to get the final FF curve and its 95% confidence interval. The results are given in the next section.

5 Results

5.1 Flood frequency relations

In this section the FFA results (Fig. 6) of the following data sets are presented:

- 1901 data set; measured discharges covering the period 1901-2018.
- 1772 data set; as above and extended with the data set of Toonen (2015), representing the systematic data set and covering the period 1772-2018.
- 1317 data set; as above and extended with 12 reconstructed historic discharges and the bootstrap resampling method to create a continuous discharge series covering the period 1317-2018.

If the data set of measured discharges is extended, we find a large reduction in the confidence interval of the FF curve (Fig. 6 and Table 2). Only extending the data set with the data of Toonen (2015) reduced this confidence interval with 5,200 m^3/s for the floods with a return period of 1,250 years (Table 2). Adding the reconstructed historic flood events in combination with

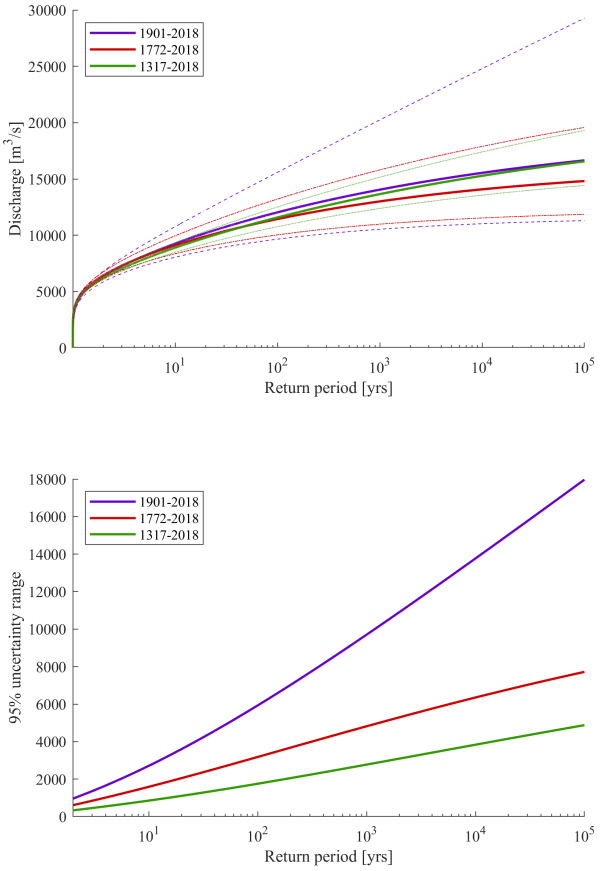


Figure 6. Fitted GEV curves and their 95% confidence intervals of the 1901, 1772 and 1317 data sets

a bootstrap method to create a continuous data set, results in an even larger reduction in the confidence interval of 7,400 m^3/s compared to the results of the 1901 data set. For the discharges with a return period of 100,000 years, we find an even larger reduction in the confidence intervals (Table 2).

Furthermore, we find that using only the 1901 data set results in larger design discharges compared to the two extended data sets. This is in line with the work of Toonen (2015). Surprisingly however, we find that the 1772 data set predicts the lowest discharges for return periods $>$ 100 years (Table 2), while we would expect that the 1317 data set predicts the lowest values according to the findings of Toonen (2015). The relatively low positioning of the FF curve constructed with the 1772 data, compared to our other 1317 and 1901 data sets, might be explained by the fact that the data of Toonen (2015) covering the period 1772-1900 has not been normalized. This period has a relative high flood intensity (Fig. 1). However, only two flood events exceeded 10,000 m^3/s. A lot of dike reinforcements along the Lower Rhine were executed during the last century. Therefore, it is likely that before the 20th century, flood events with a maximum discharge exceeding 10,000 m^3/s resulted in

dike breaches and overflow upstream of Lobith. As a result, the maximum discharge of such an event decreased significantly. Although Toonen (2015) mentions that the effect of recent changes in the river system on discharges of extreme floods of the Lower Rhine is small, we argue that it does influence the flood events with maximum discharges slightly lower than the current main channel and floodplains capacity. Currently, larger floods are possible to flow in downstream direction without the occurrence of inundations compared to the 19th century. Therefore, it is most likely that the 1772-1900 data set of Toonen (2015) underestimates the flooding regime of that specific time period influencing the shape of the FF curve.

5.2 Hypothetical future extreme flood event

After the 1993 and 1995 flood events of the Rhine river, the FF relation used in Dutch water policy was recalculated taking into account the discharges of these events. All return periods were adjusted. The design discharges with a return period of 1,250 years, which was the most important return period at that time, increased with 1,000 m^3/s (Parmet et al., 2001). Such an increase in the design discharge requires more investments in dike infrastructure and floodplain measures to re-establish the safety levels. Parkes and Demeritt (2016) found similar results for the river Eden, UK. They showed that the inclusion of the 2015 flood event had a significant effect on the upper tail of the FF curve, even though their data set was extended from 1967 to 1800 by adding 21 reconstructed historic events to the data set of measured data. Schendel and Thongwichian (2017) argues that if the flood frequency relation alters after a recent flood, and if this change can be ambiguously attributed to this event, the data set of measured discharges must expanded since otherwise the FF results will be upward biased. Based on their considerations, it is interesting to see how adding a single extreme flood event influences the results of our method.

Both the 1317 and 1901 data sets are extended from 2018 to 2019 with a hypothesized flood in 2019. We assume that in 2019 a flood event has occurred that equals the largest measured discharge so far. This corresponds with the 1926 flood event (Fig. 1), having a maximum discharge of 12,600 m^3/s. No uncertainty of this event is included in the analysis. Fig. 7 shows that the FF curve based on the 1901 data set changes significantly as a result of this hypothesized 2019 flood. We calculate an increase in the discharge corresponding with a return period of 100,000 years of 1,280 m^3/s. Contrarily, the 2019 flood has almost no effect on the extended 1317 data set. The discharge corresponding to a return period of 100,000 years only increased slightly with 180 m^3/s. Therefore, we conclude that the extended data set is more robust to changes in FF relations as a result of future flood events. Hence, we expect that the changes in FF relations after the occurrence of the 1993 and 1995 flood events would be less severe if the analysis was performed with an extended data set as presented in this study. Consequently, decision makers might have taken a different decision since less investments were required to cope with the new flood safety standards. Therefore, we recommend to use historical information about the occurrence of flood events in future flood safety assessments.

6 Discussion

We developed an efficient bootstrap method to include historic flood events in an FFA. We used a 1D-2D coupled hydraulic model to normalize the data set of Meurs (2006) for modern topography. An advantage of the proposed method is that any kind of historical information (e.g. flood marks, sediment depositions) can be used to extend the data set of annual maximum

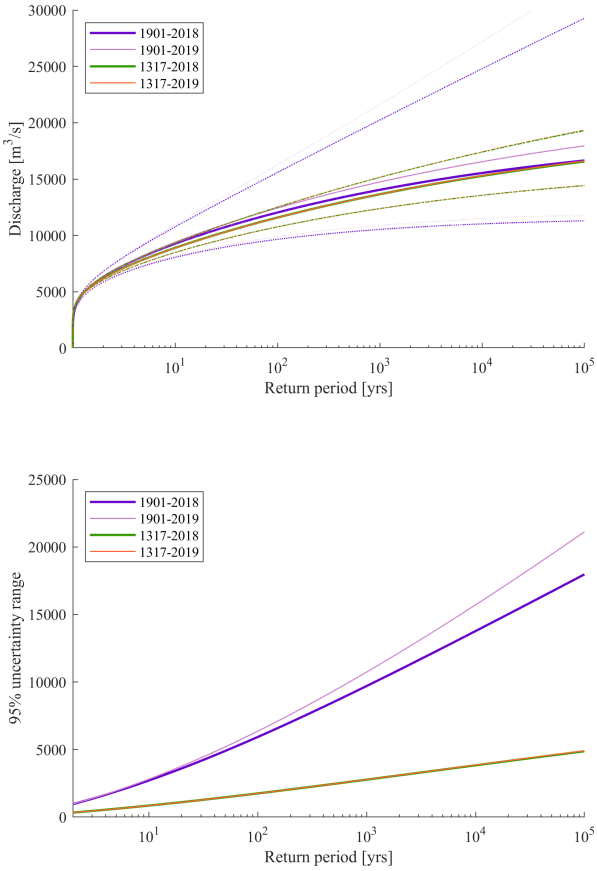


Figure 7. Fitted GEV curves and their 95% confidence intervals of the 1901 and 1317 data sets if they are extended with a future flood event

discharges as long as the information can be translated into discharges. Another great advantage of the proposed method is the computational time to create the continuous data sets and to fit the GEV distributions. The entire process is completed within several minutes. Furthermore, it is easy to update the analysis if more historical information about flood events becomes available. However, the method is based on various assumptions and has some drawbacks. These assumptions and drawbacks are discussed below.

6.1 Added value of normalized historic flood events

The results have shown that extending the systematic data set with normalized historic flood events can significantly reduce the confidence intervals of the FF curves. This is in line with the work of O'Connell et al. (2002) who claim that the length of the instrumental record is the single most important factor influencing uncertainties in flood frequency relations. However, reconstructing historic floods is time consuming, especially if these floods are normalized with a hydraulic model. Therefore,

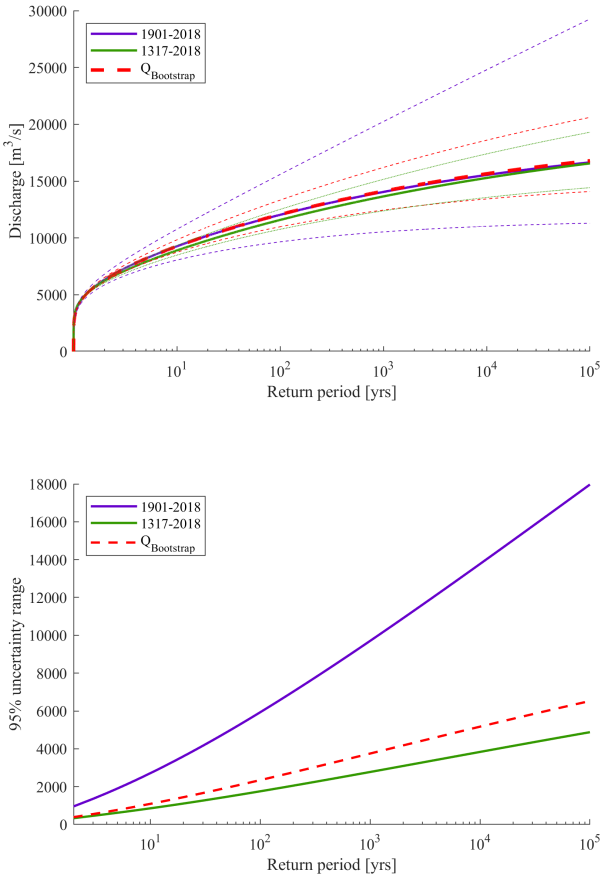


Figure 8. Fitted GEV curves of the 1901, 1317 and $Q_{Bootstrap}$ data sets

the question arises whether it is required to reconstruct historic floods to extend the data set of measured discharges. Another, less time consuming, option might be to solely resample the measured discharges in order to extend the length of the data set. Such a method was applied by Chbab et al. (2006) who resampled 50 years of weather data to create a data set of 50,000 years of annual maximum discharges.

To test the applicability of solely using measured discharges, we use the bootstrap method presented in Section 3. A data set of approximately 700 years (equal to the length of the 1317 data set) is created based on solely measured discharges in the period 1901-2018. The perception threshold is assumed to be equal to the lowest measured discharge such that the entire data set of measured discharges is used during the bootstrap resampling. Again, 5,000 discharge data sets are created to reach convergence in the FFA. This data is referred to as the $Q_{Bootstrap}$ data set.

We find that the use of the $Q_{Bootstrap}$ data set, based on solely resampling the measured discharges of the 1901 data set, results in lower uncertainties of the FF curve compared to the 1901 data set (Fig. 8). This is because the length of the measured data

set is increased through the resampling method. Although the confidence interval decreases after resampling, the confidence interval of the $Q_{Bootstrap}$ data set is still larger compared to the 1317 data set, including the normalized historic flood events (Fig. 8). This is because the variance of the $Q_{Bootstrap}$ data set, which is equal to 4.19 x 10^6 m^3/s, is still larger than the variance of the 1317 data set. For the $Q_{Bootstrap}$ data set, the entire measured data set (1901-2018) is used for resampling, while for the 1317 data set only the discharges below a certain threshold in the systematic time period (1772-2018) are used for resampling. 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. Therefore, the variance of the 1317 data set is relatively low (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 intervals are smaller compared to the $Q_{Bootstrap}$ data set.

Furthermore, the FF curve of the $Q_{Bootstrap}$ data set is only based on a relatively short data set of measured discharges and hence only based on the climate conditions of this period. Extending the data set with historic flood events gives a better representation of the long-term climatic variability in flood events since these events only have been normalized for changes in the river system and thus still capture the climate signal. We conclude that reconstructing historic events, even if their uncertainty is large, is worth the effort since it reduces the uncertainty intervals of design discharges corresponding to rare flood events which is crucial for flood protection policy-making.

6.2 Resampling systematic data set

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. Furthermore, by resampling the measured data set, we assume that the flood series consist of independent and identically distributed random variables. This might not be the case if climate variability plays a significant roll in the considered time period resulting in a period of e.g. extreme low or high flows. However, up till now no consistent large-scale climate change signal in observed flood magnitudes has been identified (Blöschl et al., 2017).

In Section 5, we found that extending the data set from 1901 to 1772 resulted in a shift in downward direction of the FF curve. This is because in the period 1772-1900, a relatively small number of floods exceeded a discharge larger than 10,000 m^3/s. Since no large flood events were present in the period 1772-1900, this data set has a lower variance compared to the 1901 data set. Using both the 1772 and 1901 data sets for resampling purposes influences the uncertainty of the FF curve. To identify this effect, we compared the results if solely the measured discharges (1901-2018) are used for resampling purposes and if the entire systematic data set (1772-2018) period is used. We find that using the entire systematic data set results in a reduction in the 95% confidence intervals compared to the situation in which solely the measured discharges are used caused by the lower variance in the period 1772-1900. However, the reduction is at maximum 12% for the return period of 100,000 years. Although the lower variance in the 1772-1900 data set might be explained by the fact that these discharges are not normalized, the lower variance may also be caused by the natural variability in climate.

6.3 Distribution function and goodness-of-fit test

In Section 5, only the results for a GEV distribution were presented. We found that the uncertainty interval of the flood event with a return period of 100,000 years was reduced with 73% by extending the data set of approximately 120 years of annual maximum discharges to a data set with a length of 700 years. Performing the analysis with other distributions yield similar results. A reduction of 60% is found for the Gumbel distribution and a reduction of 76% for the Weibull distribution. This shows that, although the uncertainty intervals depend on the probability distribution function used, the general conclusion of reduction in uncertainty of the fitted FF curve holds.

However, by only considering a single distribution function in the analysis, model uncertainty is neglected. One approach to manage this uncertainty is to create a composite distribution of several distributions each allocated a weighting based on how well it fits the available data (Apel et al., 2008). Furthermore, the uncertainty related to the use of various goodness-of-fit tests was neglected since only the Maximum Likelihood function was used to fit the sample data to the distribution function. Using a composite distribution and multiple goodness-of-fit tests will result in an increase in the uncertainties of FF curves.

6.4 Length of extended data set and considered perception threshold

The measured data set starting at 1901 was extended to 1317. However, the extended data set still has limited length compared to the maximum return period of 100,000 years considered in Dutch water policy. Preferably, we would like to have a data set with at least the same length as the maximum safety level considered such that extrapolation in FFAs is not required anymore. However, the proposed method is a large step to decrease uncertainty.

Furthermore, the systematic data set was used to create a continuous data set using a bootstrap approach. However, preferably we would like to have a historical continuous record since now the low flows are biased on climate conditions of the last 250 years. Using this data set for resampling influences the uncertainty intervals of the FF curves. If the historical climate conditions highly deviated from the current climate conditions, this approach does not produce a reliable result. In addition, the perception threshold influences the variance of the considered data set and hence the uncertainty of the FF curve. Using a smaller threshold results in an increase in the variance of the data set and hence to an increase in the uncertainty intervals. The proposed assumption related to the perception threshold can only be used if there is enough confidence that the smallest known flood event in the historical time is indeed the actual smallest flood event that occurred in the considered time period.

6.5 Comparison with Bayesian statistics

The FFA was performed based on frequentist statistics. The Maximum Likelihood function was used to fit the parameters of the GEV distribution function. However, only point estimates are computed. To enable uncertainty predictions of the GEV parameter estimates, the maximum likelihood estimator assumes symmetric confidence intervals. This may result in an incorrect estimation of the uncertainty which is specifically a problem for small sample sizes. For large sample sizes, maximum likelihood estimators become unbiased minimum variance estimators with approximate normal distributions. Contrarily, Bayesian statistics provide the entire posterior distributions of the parameter estimates and thus no assumptions have to be made. How-

ever, a disadvantage of the Bayesian statistics is that the results are influenced by the priors describing the distributions of the parameters (Neppel et al., 2010). For future work, we recommend to study how uncertainty estimates differ between the proposed bootstrap method and a method which relies on Bayesian statistics such as Reis and Stedinger (2005).

Moreover, a disadvantage of the proposed bootstrap approach is that, by resampling the systematic data set to fill the gaps in the historical time period, the shape of the flood frequency curve is influenced in the domain corresponding to events with small return periods (i.e. up to $\sim$ 100 years corresponding with the length of the 1901 data set). Methods presented by e.g. Reis and Stedinger (2005) and Wang (1990) use historical information solely to improve the estimation of the tail of the FF curves, while the systematic part of the curve stays untouched. Table 2 shows the discharges corresponding with a return period of 100 years for both the 1901 data set and the extended 1317 data set following the bootstrap method described in Section 3. We find that this discharge decreases from 12,075 m^3/s to 11,628 m^3/s by extending the systematic data set. This decrease in design discharge with 3.7% indicates that resampling the systematic data set over the historical time period only has a little effect on the shape of the flood frequency curve corresponding with small return periods justifying the use of the bootstrap method.

7 Conclusions

Design discharges are commonly determined with the use of flood frequency analyses (FFA) in which measured discharges are used to fit a probability distribution function. However, discharge measurements have been performed only for the last 50-100 years. This relatively short data set of measured discharges results in large uncertainties in the prediction of design discharges corresponding to rare events. Therefore, this study presents an efficient bootstrap method to include historic flood events in an FFA. The proposed method is efficient in terms of computational time and set-up. Additionally, the basic principles of the traditional FFA remain unchanged.

The proposed bootstrap method was applied to the discharge series at Lobith. The systematic data set covering the period 1772-2018 was extended with 12 historic flood events. The historic flood events reconstructed by Meurs (2006) had a large uncertainty range, especially for the most extreme flood events. The use of a 1D-2D coupled model reduced this uncertainty range of the maximum discharge at Lobith for most flood events as a result of the overflow patterns and dike breaches along the Lower Rhine. The inclusion of these historic flood events in combination with a bootstrap method to create a continuous data set, resulted in a decrease in the 95% uncertainty interval of 72% for the discharges at Lobith corresponding to a return period of 100,000 years. Adding historical information about rare events with a large uncertainty range in combination with a bootstrap method has thus the potential to significantly decrease the confidence interval of design discharges of extreme events.

Since correct prediction of flood frequency relations with little uncertainty is of high importance for future national flood protection programs, we recommend to use historical information in FFA. Additionally, extending the data set with historic events makes the flood frequency relation less sensitive to future flood events. Finally, we highlight that the proposed method to include historical discharges into a traditional FFA can be easily implemented in flood safety assessments because of its simple nature in terms of mathematical computations as well as of its computational efforts.

Acknowledgements. This research is supported by the Netherlands Organisation for Scientific Research (NWO, project 14506) which is partly funded by the Ministry of Economic Affairs and Climate Policy. Furthermore, the research is supported by the Ministry of Infrastructure and Water Management and Deltares. This research has benefited from cooperation within the network of the Netherlands Centre for River studies NCR (www.ncr-web.org).

The authors would like to thank the Dutch Ministry of Infrastructure and Water Management, Prof. Dr. Herget (University of Bonn) and Dr. Toonen (KU Leuven) for providing the data. Furthermore, the authors would like to thank Dr. Toonen (KU Leuven) for his valuable suggestions that improved the manuscript. In addition, the authors would like to thank Dr. Elena Volpi (Roma Tre University) and the two anonymous reviewer for their suggestions during the discussion period, which greatly improved the quality of the paper. Finally, the authors would like to thank Van der Meulen Msc, Dr. Cohen and Prof. Dr. Middelkoop from Utrecht University for their cooperation in the NWO Project Floods of the past—Design for the future.

References

Apel, H., Merz, B., and Thieken, A. H.: Quantification of uncertainties in flood risk assessments, International Journal of River Basin Management, 6, 149–162, https://doi.org/10.1080/15715124.2008.9635344, 2008.

Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B., Aronica, G. T., Bilibashi, A., Bonacci, O., Borga, M., Ivan, Č., Castellarin, A., Chirico, G. B., Claps, P., Fiala, K., Frolova, N., Gordachova, L., Gul, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirnuinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Rogger, M., Salinas, J. L., Sauquet, E., Sraj, M., Szolgay, J., Viglione, A., Volpi, E., Wilson, D., Zaimi, K., and Zivkovic, N.: Changing climate shifts timing of European floods, Science, 357, 588–590, https://doi.org/10.1126/science.aan2506, 2017.

Bomers, A., Schielen, R. M. J., and Hulscher, S. J. M. H.: The effect of dike breaches on downstream discharge partitioning, in: Abstract from NCR Days 2019: Land of Rivers, p. 28, Utrecht, The Netherlands, 2019a.

Bomers, A., Schielen, R. M. J., and Hulscher, S. J. M. H.: Consequences of dike breaches and dike overflow in a bifurcating river system, Natural Hazards, https://doi.org/10.1007/s11069-019-03643-y, 2019b.

Bomers, A., Schielen, R. M. J., and Hulscher, S. J. M. H.: The severe 1374 Rhine river flood event in present times, in: 38th IAHR World Congres, Panama City, Panama, 2019c.

Bronstert, A., Bardossy, A., Bismuth, C., Buiteveld, H., Disse, M., Engel, H., Fritsch, U., Hundecha, Y., Lammersen, R., Niehoff, D., and Ritter, N.: Multi-scale modelling of land-use change and river training effects on floods in the Rhine basin, River research and applications, 23, 1102–1125, https://doi.org/10.1002/rra.1036, 2007.

Brunner, G. W.: HEC-RAS, River Analysis System Hydraulic Reference Manual, Version 5.0, Tech. Rep. February, US Army Corp of Engineers, Hydrologic Engineering Center (HEC), Davis, USA, 2016.

Chbab, E. H., Buiteveld, H., and Diermanse, F.: Estimating Exceedance Frequencies of Extreme River Discharges Using Statistical Methods and Physically Based Approach, Osterreichse Wasser- und Abfallwirtschaft, 58, 35–43, 2006.

Frances, F., Salas, J. D., and Boes, D. C.: Flood frequency analysis with systematic and historical or paleoflood data based on the two-parameter general extreme value models, Water Resources Research, 30, 1653–1664, https://doi.org/10.1029/94WR00154, 1994.

Gaume, E., Gaál, L., Viglione, A., Szolgay, J., Kohnová, S., and Blöschl, G.: Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites, Journal of Hydrology, 394, 101–117, https://doi.org/10.1016/j.jhydrol.2010.01.008, 2010.

Haberlandt, U. and Radtke, I.: Hydrological model calibration for derived flood frequency analysis using stochastic rainfall and probability distributions of peak flows, Hydrology and Earth System Sciences, 18, 353–365, https://doi.org/10.5194/hess-18-353-2014, 2014.

Hall, J. W., Dawson, R. J., Sayers, P. B., Rosu, C., Chatterton, J. B., and Deakin, R.: A methodology for national-scale flood risk assessment, Proceedings of the Institution of Civil Engineers - Water and Maritime Engineering, 156, 235–247, https://doi.org/10.1680/wame.2003.156.3.235, http://www.icevirtuallibrary.com/doi/10.1680/wame.2003.156.3.235, 2003.

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

Herget, J. and Meurs, H.: Reconstructing peak discharges for historic flood levels in the city of Cologne, Germany, Global and Planetary Change, 70, 108–116, https://doi.org/10.1016/j.gloplacha.2009.11.011, 2010.

Hirsch, R. M. and Stedinger, J. R.: Plotting positions for historical floods and their precision, Water Resources Research, 23, 715–727, https://doi.org/10.1029/WR023i004p00715, 1987.

Klemeš, V.: Dilettantism in hydrology: Transition or destiny?, Water Resources Research, 22, 177–188, https://doi.org/10.1029/WR022i09Sp0177S, 1986.

MacDonald, N., Kjeldsen, T. R., Prosdocimi, I., and Sangster, H.: Reassessing flood frequency for the Sussex Ouse, Lewes: The inclusion of historical flood information since AD 1650, Natural Hazards and Earth System Sciences, 14, 2817–2828, https://doi.org/10.5194/nhess-14-2817-2014, 2014.

Meurs, H.: Bestimmung der Spitzenabflusse historischer Hochwasser in Koln, Diploma thesis, University of Bonn, 2006.

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

O'Connell, D. R. H., Ostenaa, D. A., Levish, D. R., and Klinger, R. E.: Bayesian flood frequency analysis with paleohydrologic bound data, Water Resources Research, 38, 1058–1071, https://doi.org/10.1029/200WR000028, 2002.

Parkes, B. and Demeritt, D.: Defining the hundred year flood: A Bayesian approach for using historic data to reduce uncertainty in flood frequency estimates, Journal of Hydrology, 540, 1189–1208, https://doi.org/10.1016/j.jhydrol.2016.07.025, 2016.

Parmet, B., van de Langemheen, W., Chbab, E., Kwadijk, J., Diermanse, F., and Klopstra, D.: Analyse van de maatgevende afvoer van de Rijn te Lobith, Tech. rep., RIZA, Arnhem, The Netherlands, 2001.

Reis, D. S. and Stedinger, J. R.: Bayesian MCMC flood frequency analysis with historical information, Journal of Hydrology, 313, 97–116, https://doi.org/10.1016/j.jhydrol.2005.02.028, 2005.

Sartor, J., Zimmer, K. H., and Busch, N.: Historische Hochwasserereignisse der deutschen Mosel, Wasser Abfall, 10, 46–51, 2010.

Schendel, T. and Thongwichian, R.: Considering historical flood events in flood frequency analysis: Is it worth the effort?, Advances in Water Resources, 105, 144–153, https://doi.org/10.1016/j.advwatres.2017.05.002, 2017.

Stendinger, J. R. and Cohn, R. A.: Flood frequency analysis with historical and paleoflood information, Water Resources Research, 22, 785–793, 1987.

Te Linde, A. H., Aerts, J. C., Bakker, A. M., and Kwadijk, J. C.: Simulating low-probability peak discharges for the Rhine basin using resampled climate modeling data, Water Resources Research, 46, 1–19, https://doi.org/10.1029/2009WR007707, 2010.

Tijssen, A.: Herberekening werklijn Rijn in het kader van WTI2011, Tech. rep., Deltares, Delft, the Netherlands, 2009.

Toonen, W. H. J.: Flood frequency analysis and discussion of non-stationarity of the Lower Rhine flooding regime (AD 1350-2011): Using discharge data, water level measurements, and historical records, Journal of Hydrology, 528, 490–502, https://doi.org/10.1016/j.jhydrol.2015.06.014, 2015.

Van Alphen, J.: The Delta Programme and updated flood risk management policies in the Netherlands, Journal of Flood Risk Management, 9, 310–319, https://doi.org/10.1111/jfr3.12183, 2016.

Van der Most, H., De Bruijn, K. M., and Wagenaar, D.: New Risk-Based Standards for Flood Protection in the Netherlands, in: 6th International Conference on Flood Management, pp. 1–9, https://doi.org/10.1017/CBO9781107415324.004, 2014.

Van Hal, L.: Hydraulische randvoorwaarden Rijn en Maas, Tech. rep., RIZA. Memo RYN2003-12(A), Arnhem, The Netherlands, 2003.

Verheij, H. J. and Van der Knaap, F. C. M.: Modification breach growth model in HIS-OM. H.J. Verheij, Tech. rep., WL | Delft Hydraylics, Delft, The Netherlands, 2003.

Vorogushyn, S. and Merz, B.: Flood trends along the Rhine: The role of river training, Hydrology and Earth System Sciences, 17, 3871–3884, https://doi.org/10.5194/hess-17-3871-2013, 2013.

Wang, Q. J.: 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, Journal of Hydrology, 120, 115–124, https://doi.org/10.1016/0022-1694(90)90145-N, 1990.