[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 will provide more knowledge about how the discharges at Lobith were reconstructed by Toonen (2015) as well as the reconstructions at Cologne performed by Herget and Meurs (2010). Furthermore, the following table will be 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 3 lines 13-14 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 will be added with in green te new text:

Regarding the 1901-2008 data set:

"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:

"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:

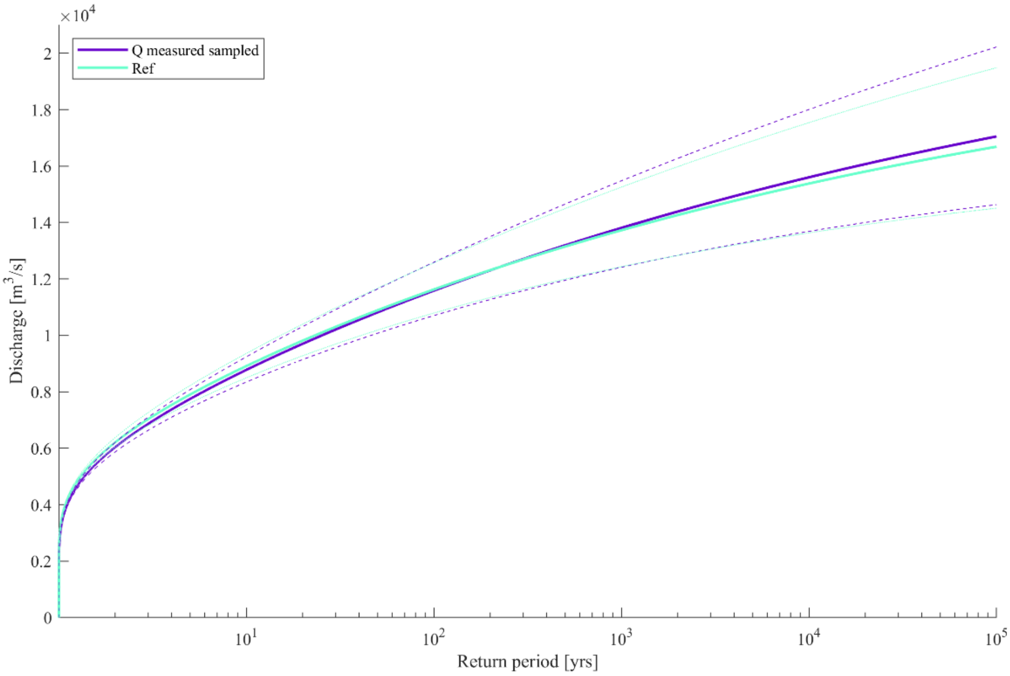
"In this study, the 1D-2D coupled modelling approach as described by Bomers et al. (2019) is used to normalize the data set of Meurs (2006). This normalization is performed by routing the reconstructed historical discharges at Cologne over modern topography to estimate the maximum discharges at Lobith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

However, do note that the lower variance in the 1772-1900 period compared to the 1901-2018 period is 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?

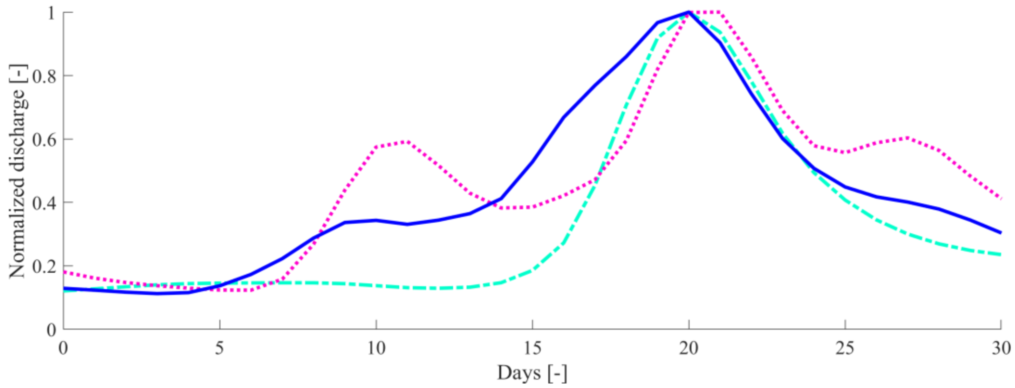
Please note that de 1D-2D coupled model is only based on the current geometry and current dike strengths. This is because only then normalization can be performed. So, whether dike breaches occurred during the historical flood events between Andernach and Lobith may be interesting from historical point 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 will provide more insights in the 1D-2D coupled modelling approach in the revised manuscript and particularly about the dike breach parameters.

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 will be added in the revised manuscript.

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 will provide more information about the upstream discharge wave shapes in the revised manuscript.



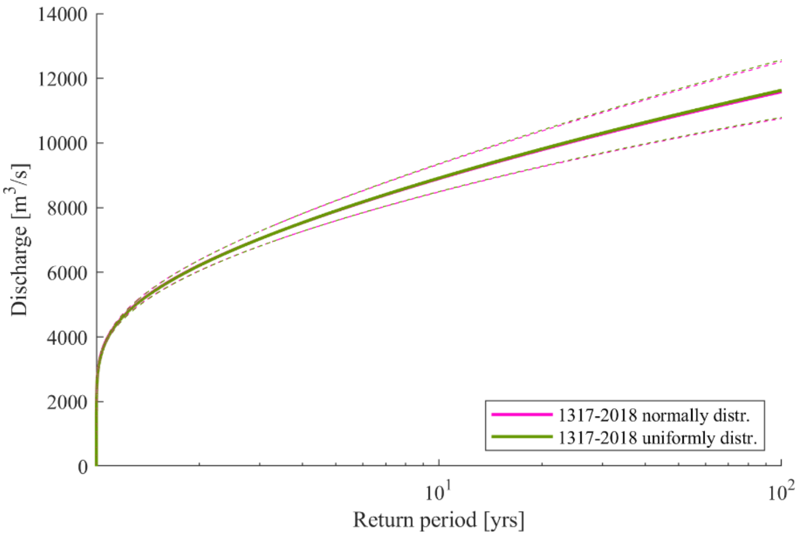
Furthermore, it would be interesting to know if any of 12 historical flood events where winter events, where ice draft/ice jams could/did play a role.

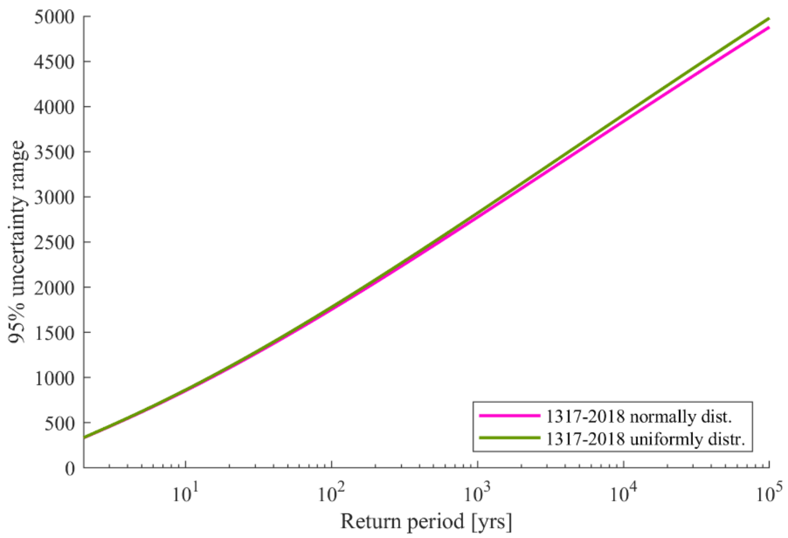
All flood events were winter events, except for the flood event in 1342 that took place in July. However, the flood events caused by ice jams were excluded from the analysis by Herget and Meurs (2010) because of the different hydraulic conditions. All flood events considered are thus caused by high rainfall intensities. This will be added to the revised manuscript.

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.





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. 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 will be 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.

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 will rewrite the paragraph in the revised manuscript. Furthermore, we will rewrite the paragraph in a more discussion style and replace this section towards the discussion section.

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 will move 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.

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. The sentence will be rewritten in the revised manuscript.

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 will include this uncertainty for all ACDP-measurements. 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 will rewrite the text such that this becomes clearer.

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

Please see above.

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 will be moved to the results. It is indeed true that we did not include the uncertainty caused by different distributions and parameter estimation methods. We will remove the statement from the manuscript and add in the discussion the effects of using a combination of different distributions on the uncertainty intervals.

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 figures will be replaced by the following figure such that now the national boundaries, scale bar, north arrow, names and model boundary are given.

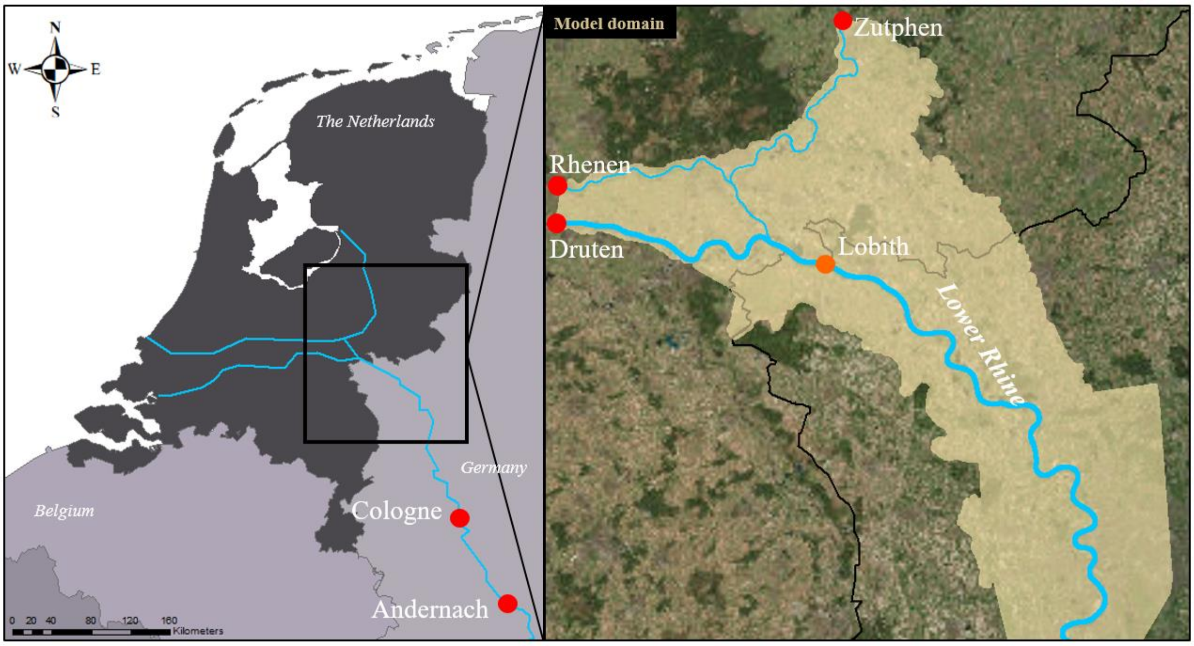
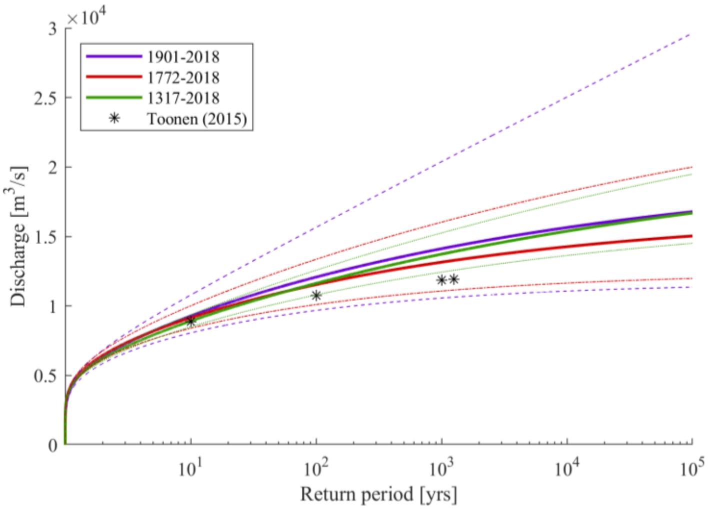


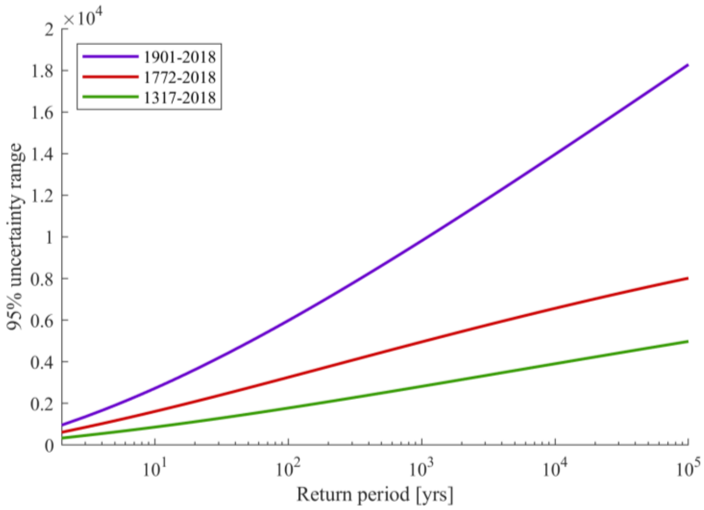
Table 1: The results of Toonen (2015) can be omitted in this table from my point of view.

The results of Toonen (2015) will be 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 will be 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 will be 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