Point-by-point response to Reviewer #1

Florian Ehmele on behalf of all co-authors

November 11, 2021

Dear Reviewer No. 1, Thank you very much for your work and the useful and valuable comments that will help to improve the scientific quality of our manuscript. Below you will find your comments given in gray and our responses to the individual points in black. Please also consider our comments to Reviewer 2 as there is some coincidence of the comments and the corresponding answers.

This is an interesting paper that describes the use of a large ensemble of regionally downscaled multi-GCM forcings to drive a hydrological model for impact assessments. The issue of long return period extremes is highly relevant. The paper is very well written, clearly structured and to the point. However, there are some unfortunate shortcuts regarding the model validation which needs to be handled differently.

Thank you very much. We hope to implement your comments in a sufficient way.

Main comments:

Both the bias correction and the HBV set ups are validated on the calibration period. While I can accept this for the bias adjustment because it is not anywhere applied outside of the calibration period, it is a big issue for the justification of the hydrological model. HBV is currently calibrated and validated on the same period (1961-2006) based on precipitation and temperatur forcing from gridded observational data sets. When validated on that same period, the results are very good, as seen from the very high NSE values. However, we still know nothing about the model's performance on data it has never seen before, and the main results are based on the downscaled model data. I urge the authors to at least perform a split sample validation where calibration and validation periods are independent, or even a cross-validation. This is standard practice in hydrological model validation.

We agree with the reviewer that proper practice is a split between calibration and validation period in case of the HBV model. For the revised version we will split the considered time period from 1961 to 2006 into a calibration and a validation phase and present the NSE results for the new validation period only.

Bias correction is only performed for precipitation, and no information about potential bias in temperature and how it might affect results is provided. Because temperature, and its translation into evapotranspiration, is an important input to the water balance of the model, it should not be neglected. I would like to at least see a justification for why temperature is not bias corrected (being that the bias is low). In some cases it can be neglected for certain extremes where the pre-conditioning of the river is of minor importance, but also that needs some additional analysis and commenting in the text.

We agree that evapotranspiration and therefore also temperature is important for the total water budget. The LAERTES-EU temperature data have also been bias-corrected using the quantile mapping approach with a Gaussian distribution function. The bias-corrected temperature data have been used in line with bias-corrected precipitation. Nevertheless, the dominant factor in case of the major flooding events is precipitation so we focus on the precipitation part of the bias correction. We will add a comment on that.

The concluding main result of the paper is presented in figure 7. Although the result is compelling and seemingly clear, the details may occlude the actual results. First, the

lenght of each timeseries has a large effect on the GEV fits and their robustness, as argued in the introduction. Please add the record lenght, i.e. the number of years, in the legend for each data set.

We agree with the reviewer that the length of the time series has a crucial impact on the estimated distributions. As requested we will add the length of each time series in the legend of the plot and in terms of discharge data also in Table 2.

Second, it would help the reader a lot to also see the confidence intervals. With so many lines, it might get too busy, but I think adding e.g. the confidence interval for the "Q obs - Weibul" and "LAERTES-EU BC" would be very informative. The confidence intervals would convey two results, one is the fair comparison of the observations and the model that would show the observations results essentially useless beyond 50 years (depending on the lenght of the timeseries), and the other is the added value of the multi-realization simulations which add statistical robustness for the longer return periods.

Thank you for this comment which we totally agree with. For short time series it is necessary to fit distributions to the data to extrapolate discharge at higher return periods and the extrapolation has significant uncertainty. But for LAERTES-EU we have over 12000 years and need only discharge at RP 2000 for the application of this project. This can be read visually from the empirical graph, and hence theoretical distribution isn't needed for extrapolation. That is also why the LAERTES-EU lines are curvy. We plotted the sorted modeled discharges (empirical distribution), not a smooth parametric distribution. There is some uncertainty in assigning return periods to the sorted yearly maximum discharges (method called plotting positions) but it is tiny compared to uncertainty of extrapolated data. For a profound estimate of the confidence intervals of an empirical distribution, the dataset has to be split into sub-samples with each subset being long enough for a robust estimation of high return periods up to 2000 years. This would be possible only for a much larger dataset than LAERTES-EU. Nevertheless, we will add the confidence intervals for the parametric distribution estimates of Q obs Weibull and HYRAS in Fig.7 and S7-S11. We will also adjust the text on that.

Minor comments:
L69: Please clarify what you mean with "isolate the effects".

It was meant to elaborate the changes that the bias correction brings to both precipitation and discharge statistics and the added value for an application such as the one presented in this study. We will rewrite this sentence for clarification.

L85-94: Please describe more details about the LEARTES-EU multi-model. It is currently not clear what the driving GCMs are; especially that they area mixture of assimilated reanalysis, decadal initialized forecasts and free GCM simulations. Please repeat more from Ehmele et al. (2020) which provides a good summary, enough for the reader to understand from this paper alone.

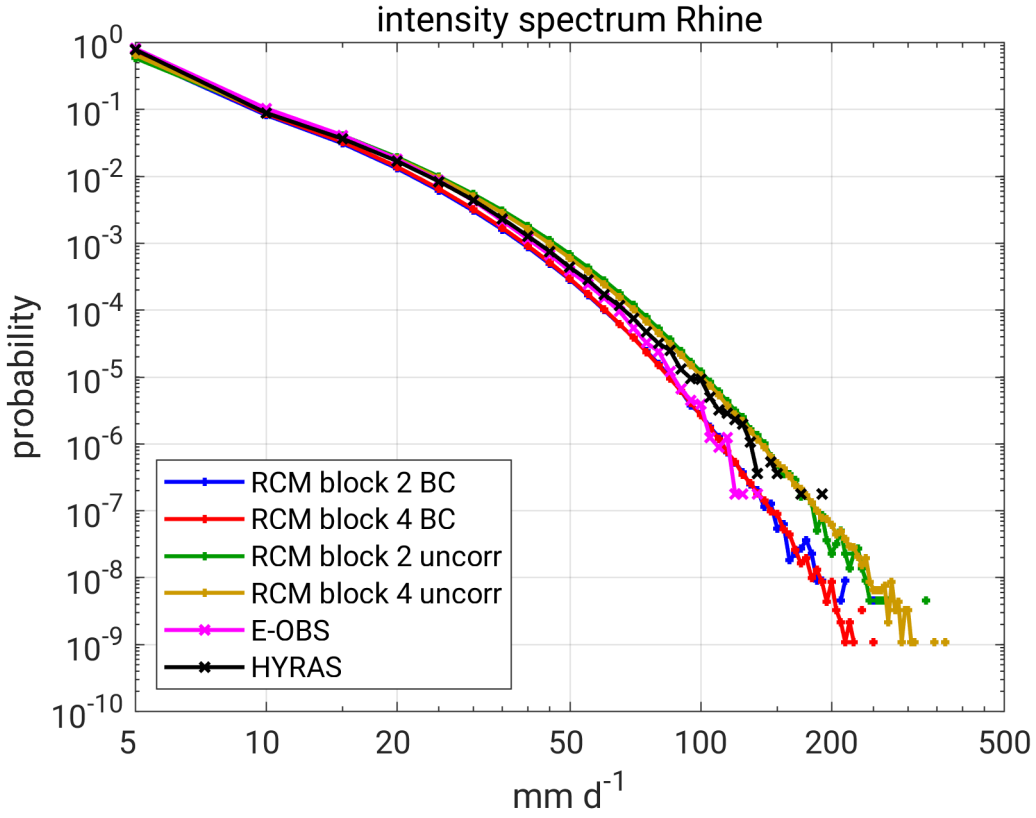
We will rewrite the section on the LAERTES-EU dataset and provide more information similar to the original study by Ehmele et al. (2020) but still keeping it short and concise.

L176, 185: I would avoid describing differences between data sets as "bias", but rather use the word "difference" unless you include a well established ground truth observational reference.

We agree that "bias" is misleading in this context. As suggested we will change to "differences" where appropriate.

Figure 3: Please consider using a log-log scale, which would better show differences between the data sets for the (0,100) mm/day range.

Thank you for this suggestion. We already tried a log-log scale plotting, which is added below. We agree, that in theory the range between 0 and 100 mm can be better recognized in a log-log scale. As shown in the figure, there are only small differences in this intensity range below 50 mm which can also be recognized in the single y-log version. Furthermore, the more interesting and relevant part of the distribution is the heavy tail which is better represented in the single y-log scaling. So we decided to take Figure 3 as it is.



L307: "different forcing and/or assimilation schemes". I refer back to my earlier comment that the LAERTES-EU sources needs to be better described.

We will add more information on LAERTES-EU, please refer to the comment above for details.

L310: "consistend data for precipitation and temperature". This is not really true after bias correction. The depencency between the variables can be severely impacted. You have also not described the potential temperature bias and how it migh affect rain/snow distribution and timing over the year. It might not be useful to retain the dependence it is errouneous?

We agree that "consistent" is potentially misleading in this context and care has to be taken when using it. As mentioned before the temperature bias is small and the dependency disruption is limited so that the data set can be treated as almost consistent or "consistent to a large degree". Furthermore, the large-scale dynamics that produce specific weather patterns and precipitation fields are not influenced by the bias correction so that a synoptic situation leading to heavy precipitation remains the same after bias correction. We will change the wording in the text accordingly.

Figure S7-11: Please change "Observed - Weibul" to "Q obs. - Weibul" as in the main text figure.

This was accidentally forgotten to adjust. We will fix this in the revised version.