Response to reviewer for revision of "Technical Note: Evaluation and bias correction of an observations-based global runoff dataset using streamflow observations from small tropical catchments in the Philippines" by Ibarra, David and Tolentino for HESS. March 2021.

All responses are indented and given in blue below.

Dear Dr Ibarra,

Thank you again for bearing with me while seeking a report from the original Referee #2. I am happy to let you know that I have finally received this report, and that the reviewer only has raised minor comments.

However, these comments are important enough to be addressed adequately. In particular, I agree with the reviewer that the procedure that you use to calculate the flow duration curves is not suitable. Flow data from different catchments should not be pooled because they represent different datasets with different statistical characteristics. So you should follow the reviewer's suggestion of calculating individual percentiles and then analysing the statistical characteristics of each group of percentiles, such as median/mean and range.

I appreciate that the other comments are unrelated to previously raised points, but I also concur with the reviewer that they are pertinent enough to improve the manuscript further, and should be easy enough to address in a minor revision.

I am confident that you can address these issues swiftly. I expect to be able to make a decision on a revised manuscript without the need for further exterenal review (but conditional on a further editor review).

Kind regards
Wouter Buytaert
handling editor

> We thank the editor for the chance to review the manuscript. We make the corrections suggested by the review and in addressing the reviewer comments regarding Figure 5, after trying this several different ways, we felt that the best way to visualize what we are trying to describe in the text (section 3.1) was their suggestion regarding bands of the FDC distributions, but because our dataset is not overly large, we actually plot all individual catchments and paired GRUN grid cells as flow duration curves, grouped in the 4 climate types. In describing the different parts of the plots in the text, this figure now actually enhances what we describe. Finally, thank you for handling our manuscript, we appreciate the great efforts of this second reviewer and you to improve the clarify and potential impact of this work, and hope that it is now suitable for publication in HESS.

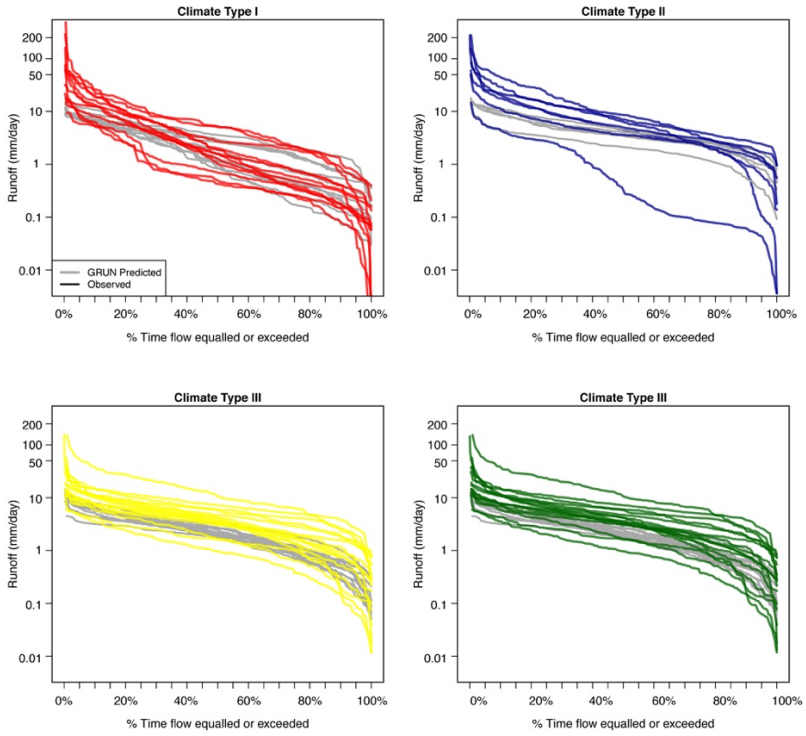
The manuscript has been significantly improved. I like the new boot strapping results but I don't think that the new analysis of the flow duration curves is entirely correct. There are some other

aspects of the manuscript that can be changed to improve the manuscript and to make it more impactful/useful for other researchers.

Thank you for the through reviews, we are glad that the new boot strapping results clarified the previous issues, and thank the reviewer for the thorough and numerous edits in their annotated PDF, these improved the clarity of the manuscript greatly.

1) I like the new addition of the comparison of the flow duration curves but I don't think that you can pool the runoff data from all the catchments and then calculate one flow duration curve from that pool of data. This automatically results in a more extreme flow duration curve (i.e., will be steeper) than the flow duration curves of the catchments in the region as it includes both the most extreme peak flow (from the most extreme catchment) and the most extreme low flow (which probably is for a different catchment). The better analysis would be to calculate the flow duration curve for each catchment individually (as you have already done) and to then calculate for each percentile the mean or median of the values of the flows from the different catchments for that percentile. Alternatively, one could plot the width of the band that spans all the different flow duration curves for the observed data and the GRUN data in a region.

This was an excellent point brought up by the reviewer, we appreciate them catching that pooling across the region did not make sense. Rather than calculate mean/median based on percentile, which would not show the full range, we have done a variation on the latter option suggested plotting bands of the FDCs (also mentioned by the reviewer as a good option in the annotated PDF), showing all the different FDC's for each individual catchment and paired GRUN grid cell on the new Figure 5 (below), this shows the same points we make in the text and we have adjusted the text and captions accordingly.



2) Table 2: I suggest to also add the range of VE and NSE values when they are calculated for the individual catchments – and perhaps the mean or median value as well. This could be useful for other researchers and helps to understand the variability in the results. I have a hard time getting my head around what for example the pooled NSE values mean and how much they are biased to some catchments with either really poorly estimated flows or catchments with overall very high runoff. Furthermore, I think that it would be useful if the metrics for the bias – corrected data were given in the same order as for the "raw" GRUN data (or perhaps in several rows below the results for the "raw" data). To understand the RMSE values, it would be useful to know the average or median monthly flows for the different climate zones, so perhaps give them in the caption.

> Great points and suggestions as this table is not yet too big. We have 1) added median and ranges as additional rows, 2) put the raw and bias corrected results in the same order (moved the bias corrected RMSE column to now be the right most column, and 3) put median monthly flow for each climate zone in Table 1 to contextualize the RMSE values. We also added the columns for bias corrected comparison to observations to Table A1 (which were then used to make the summary statistics reported here).

3) At some places it is said that the GRUN dataset is not really meant to be used to predict the discharge for individual catchments but can be used for larger scale questions, such as how much runoff reaches the ocean. This should be mentioned much earlier in the text (see suggestion in the annotated pdf). Furthermore, one of the things that is lacking is therefore a comparison of the total observed flow from all your catchments for the 10 or 20 year period with the most data and the GRUN estimates for this period and in how far the variation in the GRUN estimates of total annual flow from these catchments reflects the variation in the observed annual flow (i.e., what is the VE, r2 and NSE for total flow from all catchments combined per year for this period).

> For the first point we have adjusted based on the annotations in the PDF. The first paragraph of section 2.5 now reads: "To assess the performance of GRUN, we use a suite of metrics commonly used to assess model performance in hydrologic studies. Given the emphasis on a country scale evaluation of GRUN we primarily focus below on results in aggregate grouped by climate type or for all catchments. These metrics are calculated for each individual catchment (n=55) and in aggregate for each climate type (n=4; see below) shown in Figure 1. GRUN was not intended to be used for estimating discharge for single small catchments, therefore we focus on the aggregated data but also report the range for the result for the individual catchments."
>
> For the second point, we agree, however we actually have additional work currently underway with analysis in the works doing an extensive analysis on this dataset. In particularly we are also extended this comparison to pre-WWII discharge datasets and will look at the time component of this dataset over the 20th century as suggested. For now, however, we feel this is outside the scope of what we hoped to accomplish in this paper.

4) I don't understand why for the bias correction, first a constant is subtracted from the simulated runoff, and then a linear regression is fit through the scatter plot of the observed and simulated data and both the constant (intercept) and slope are applied to corrected the simulated data.

Would the intercept not take care of the bias that you try to fix with the subtraction of the constant first?

> This would be the case if it was done in linear space but not in log-transform space however, as described and displayed in figures 2 and 4, we do this bias correction in log-transform space (because runoff distributions are not normally distributed) thus requiring first an additive bias correction and then a linear regression (in log-transform space). This is similar to what is done in much of the climate modeling bias correction literature for observations that are non-normally distributed.

5) L265: I don't understand why averaging to monthly values would increase the bias. I would rather expect it to decrease the bias as a day with too much flow would be compensated by a day that has too little flow (regression towards the mean). I expect averaging to monthly values to lead to less extreme values and less bias than for e.g., daily values. A bit more explanation on your reasoning for this would be useful.

> Similarly, we view this as being the case because the daily or sub-daily observations are heavier tailed. We have added this explanation and note that our collaborators are exploring this detail in more detail right now associated with flood events in the Philippines. The revised sentence now reads: "Further, because of the monthly (rather than daily or sub-daily) output of GRUN and the comparison carried out here, the biases observed in our analysis could be due to averaging of the input products and/or discharge datasets to monthly values, which would mask extreme values influencing the observational dataset since the runoff distribution of daily or sub-daily runoff is heavier tailed than that of monthly distributions."

6) L275-280: This part doesn't fit well here. Merge with the text in the conclusion. But above all, try to be more realistic and don't oversell the fit for the GRUN data.

> Agreed this was deleted here and moved down as suggested in the Reviewer's PDF. We moved the two references to the conclusion as we felt it was important to keep these references in so that people can see what has been done previously in this realm.

7) More technical or editorial suggestions to further improve the text are given in the annotated pdf. Note that I don't expect a formal response to these suggestions – nor that you implement all of them.

> We implemented all of the technical and editorial suggestions. Thank you for thoroughly going over our manuscript to improve the clarity and writing!.