Response reviewers for "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. December 2020.

All responses are indented and given in green below.

Dear Dr Ibarra,

thank you for bearing with us while we awaited the referee report. We have now received this, and as you will notice, the reviewer still has substantial concerns about the manuscript. I think that these are justified, especially in view of the fact that you pitch the study as a methodological advance. I agree with that focus, but it does mean that the manuscript should present substantial methodological novelty to arrant publication. As the presented method is essentially a type of bias correction, in itself this novelty is not very high. However, I agree with the reviewer that it can be enhanced by putting a stronger focus on the implementation, validation, and discussion of the errors of the method, as the reviewer elucidates in their 5 points.

> We thank the editor for the opportunity to revise our manuscript. We have taken the time to thoroughly revise this work, including new analysis and figures, implemented changes suggested in the reviewer's comments/edits in the annotated PDF and expanded our discuss of the limitations of this work.

In addition, the reviewer also makes some very pertinent specific comments, which would also need to be addressed in an eventual revision.

> All specific comments and suggestions in the PDF and below have been implemented in this revision.

I would be willing to consider a revised version of your manuscript if you are willing to address all these issues but please note that I will seek further approval of the reviewer to ensure that their comments have been addressed to their full extent.

> Thank you for your willingness to consider this revised version of the manuscript. We sincerely hope that you feel that we have addressed the reviewer comments in a thorough manner.

Kind regards

Wouter Buytaert
handling editor

It was nice to read this manuscript again. I still think that the data are unique and that the comparison with the GRUN dataset is useful – even if to just show the errors in blindly using these model outputs. However, I still have several main comments. These include 1) the need for some more comparisons of the observations and GRUN output (e.g., flow duration curves or

flow percentiles) to strengthen the analysis, 2) the need to discuss the effects of errors in the rainfall data used in the GRUN 'simulations', 3) the mismatch between the interpretations regarding hydrological processes and the monthly time scale of the data, 4) the way the bootstrapping for the bias correction factor is implemented, and 5) the need for rewording some of the text to make it more accurate and improve readability.

> We thank the reviewer for such thorough comments and suggestions, as well as editing the language in the PDF and providing comments in the review and in the PDF. We have revised the manuscript accordingly and taken all suggestions that were given, this has included entirely overhauling the results and discussion with new analyses, a different bootstrapping methodology to report the bias correction's uncertainty, and inclusion of flow duration curves by climate type. We greatly appreciate the guidance from both the reviewer and the editor and feel that this manuscript has significantly matured over the course of peer review. With respect to the 5 key points to address, we provided responses here as well as to the more detailed responses in the subsequent pages:

1) the need for some more comparisons of the observations and GRUN output (e.g., flow duration curves or flow percentiles) to strengthen the analysis,

> This was a great suggestion. ***We have added an entirely new paragraph, new figure and a new supplemental figure*** showing flow duration curves. This comparison is nice because it allows the reader to visually see the biases, as well as underprediction of maximum runoff values by GRUN versus the observations.

2) the need to discuss the effects of errors in the rainfall data used in the GRUN 'simulations',
3) the mismatch between the interpretations regarding hydrological processes and the monthly time scale of the data,

> These two comments were addressed in tandem in the discussion of the rainfall data and the hydrological processes with respect to the monthly time scale of the data. ***See the response and substantial new text below*** associated with answering the questions regarding L246-249.

4) the way the bootstrapping for the bias correction factor is implemented, and

> This was an excellent suggestion that we somewhat blindly overlooked as we incorporated the other reviewer's suggestions during the previous iteration. ***We have now bootstrapped by catchment (not by random picking by month) and also added a figure and associated text*** to demonstrate the uncertainty of performing such a bias correction using limited data. We found that at least 10 catchments (of more than 10 years of data) would be necessary (in the case of our dataset), and in the new figure assess both the asymptotic nature of the bias correction coefficients as sample size is increased and the reduction in the associated confidence intervals.

5) the need for rewording some of the text to make it more accurate and improve readability.

All of the reviewer edits in the PDF were incorporated. Unclear phrasing (below and in the comments of the PDF) were adjusted. We also edited the manuscript for readability. We greatly appreciate the reviewer's efforts to strengthen this manuscript via thorough editing.

Specific comments:
L 100: I still do not understand the description of the data quality categories. What is meant by "the actual gauge height vs height computed"? I assume that you are looking at the rating curves here. Do you mean that xx% of the level observations are within the range of the measurements used to create the rating curve? or something else?

That is right, this is with respect to comparison to the inferred heights from the rating curves when compared to manual daily discharge measurement. Note that we do not use the daily discharge in this work, or these ratings explicitly, and thus have decided to delete this sentence commenting on the accuracy of this data, instead referring the reader the to our previous work (Tolentino et al., 2016):

> "A discussion of the accuracy of this data based on comparison to manual daily discharge measurements can be found in Tolentino et al. (2016)." (lines 97-98)

Figure 1: The font is very small – particularly when the figure is rescaled to the journal pages.

Thank you for pointing this out, we have rescaled this figure and enlarged all fonts for the labels, legend, titles etc. We hope that this figure is now much clearer.

L189-197: This section is a mixture of explanations on what is discussed and shown in the next sections and some initial results. I would include the results in section 3.1 and significantly shorten the remainder of the section or remove it completely so that you can use the "word space" to show more comparisons or more thoroughly discuss the results.

Agreed we moved the results in this short intro section into the following section and reorganized the results (see two comments below).

L190: Is a VE of 0.50 really a reasonable prediction? Doesn't it suggest an error of about 50%!! Section 3.1. Mention the range (and average) of the NSE and NSE(log) values. What are they and for how many of the catchments is it better than 0 or 0.5? This would actually tell me if GRUN has some skill in predicting the flow across a region.

Okay, we agree this is not a good prediction. For a global model we think this is "somewhat" well predicted. We have adjusted this opening line of the section to say as such:

> "Average runoff values among all catchments are somewhat well predicted by GRUN." (line 186)

Section 3.1 and 3.2: The structure of these sections is a bit confusing and leads to repetition. It is probably better/more logical to first discuss the pooled data as you do in 3.0, then the range of NSE and NSE(log) values for the individual catchments, the prediction of the average and

median flows, the prediction of the interquartile range, and then finally the prediction of the peak flows and minimum flows.
Most of the comparison of the data (and section 3.1) focuses on the peak flows. This is interesting but since this is the absolute peak it is also prone to errors in the data or just a mismatch between the GRUN and this one month with the highest flows. What about also adding a comparison of some other metrics that describe the overall peakflow fits, such as the 95th or 99th percentile of the flow or the 5-year return period monthly flow? I would have liked a comparison of the flow duration curves as well. Overall, there could have been more analyses than just the mean, max and min flow that is currently included. I think that adding a few more comparisons would strengthen the manuscript.

We agree that these sections needed further analysis and restructuring, we have done the following based on the reviewer's extensive comments in the PDF and above:

- Merged and reorganized content from 3.0, 3.1 and 3.2. As suggested by the reviewer in the PDF this is now in the order of: country wide results and average flow results, IQR results, max and min flow results, and then finally a new paragraph on the new flow duration curves (new Figure 5) by climate type.
- Flow duration curves are now included by climate type in the main text (Figure 5), comparing GRUN versus the observations, and then also for each individual catchment in a new supplementary Figure (Figure A2).

The new reworked section 3.1 is as follows (lines 189-219):

"Average runoff values among all catchments are somewhat well predicted by GRUN. Across all observations the r-squared of the correlation between GRUN prediction and observation was of 0.372 and VE of 0.363 (Table 2). Using log(runoff) values (following Criss and Winston, 2008) this improves to an r-squared of 0.546 and a volumetric efficiency (VE) of 0.733, suggesting reasonable utility in the GRUN product at a country scale for the Philippines, despite no training data from the Philippines being used in the creation of GRUN. The RMSE across the dataset was 2.648 mm/day (Table 2). NSE and NSE-log10 values, which scored overall values across the dataset of 0.091 to 0.453, respectively, ranged, from -10.70 to 0.68 and -11.53 to 0.76, respectively, for individual catchment comparisons, with median values of 0.02 and 0.24, respectively. More than half of the catchments, 29 of 55 scored NSE values greater than 0, with only 5 catchments scoring values greater than 0.5. Similarly, 32 of 55 catchments scored NSE-log10 values greater than 0, with 12 catchments scoring values greater than 0.5.

In Figure 3A median values of runoff (black dots) given a volumetric efficiency (VE) metric of 0.509 across all catchments, the average (mean) difference between the observed median and simulated values is +16%. The median and interquartile ranges (IQR, 25% to 75%) shown in Figure 2E overlap between GRUN and the observations. For five catchments of large (n=2) and relatively small sizes (n=3) the IQR of the observations does not overlap with the GRUN runoff IQR. The three small catchments are Climate Type III (yellow) and the two large catchments are Climate Type IV. In two catchments of moderate size the GRUN IQR is greater than the observed IQR runoff range. Looking at extreme monthly values (maximum and minimum) over the period of observation

demonstrates significant underprediction in wettest conditions (orange dots in Figure 3A and 3D) with almost all catchments' maximum observations falling above the 1:1 line and a lower VE score (compared to the median runoff VE score) for maximum values of 0.194. The minimum values plot around the 1:1 line and are more evenly distributed, however the VE score of 0.154 is similarly low due to greater spread than the median values.

Regardless of Climate Type, a general underestimation of the model is seen for the highest runoff months, looking at the distributions by basin. This is especially evident in Climate Types I and II with pronounced wet seasons as also shown by their lower r-squared values (Figure 4) and lower VE values. Climate Type II also has the highest RMSE value of 4.55 mm/day (compared to an average observed flow of 9.03 mm/day). Climate Types III and IV have comparable r-squared and VE values though skewness towards underprediction during the highest runoff months is still evident, particularly for Climate Type IV. These patterns are particularly evident looking at Figure 5, which shows flow duration curves (FDC) by climate type (see Figure A2 for individual catchments). Such an analysis allows for inspection of runoff distributions and biases across the range of observed and predicted values. At low flow (high exceedance probability, >80%) there is reasonable agreement in the shape and magnitude of the distributions for Climate Types I and IV between GRUN and the observations (bottom right of the FDC plots). Climate Type III demonstrates consistent bias across all runoff values less than an exceedance probability of ~90%. At high flow (low exceedance probability, <20%), as also noted above, runoff for all Climate Types is underestimated by GRUN, with Climate Types I and II showing the greatest discrepancy (top left of each FDC plot)."

L226: A RMSE of 4.55 mm/d seems very large. Please put these values into perspective. How does this compare to the average flow?

Good point, it's significant. The average flow is 9.03 mm/d for Climate Type II, we have now commented on this in a parenthetical added to this sentence:

"Climate Type II also has the highest RMSE value of 4.55 mm/day (compared to an average observed flow of 9.03 mm/day)." (lines 209-210)

L246-249: The explanation given here seems not plausible. It would be fine if we looked at hourly or daily data but here monthly data are used. It seems very unlikely that for the larger catchments (which are still not very big) the flood events last multiple months! Routing simply isn't that slow. As far as I know there are also not that many very large lakes in the Philippines that could buffer all this water for the larger catchments. Does it rather mean that small catchments are more dominated by fast flow pathways, such as ssf, and larger catchments by slower pathways, such as groundwater flow? Although I think that streambed infiltration is important in some areas, I am not sure if in such a wet country like the Philippines, there is really that much loss of water from the stream into the aquifers to delay the streamflow response by several months. I like the attempt to describe the differences in terms of hydrological processes (here and on L275-280) but think that the monthly time scale of the data aren't fully considered in these interpretations. Yes, whether runoff is generated as overland flow or subsurface

stormflow has a huge effect on the hourly or 5-min peakflows but for the monthly runoff values, this effect should be fairly small as both flow pathways will transport the water to the stream within the monthly timescale.
The larger issue is likely the rainfall. For larger catchments, the average rainfall intensity and variation in the rainfall is less (due to the averaging over larger areas) and perhaps better predicted or represented by the GSWP precipitation data that are used in GRUN. Add some discussion on what is known about the bias in the GSWP precipitation data – and bias in the variability of precipitation. There is currently no information on how any bias in precipitation for the Philippines in GSWP may have caused the huge bias in the GRUN streamflow. Considering the need for significant rescaling of the GRUN streamflow predictions. It seems that there must be a bias in the input (i.e., rainfall data) used for the streamflow predictions. Otherwise, the mass balance can't work out. I think that more discussion on this is needed.

We agree this is a very valid point we overlooked previously and is the more likely candidate. To rectify we softened the sentence in question, deleted the second sentence, and then added an addition section to this paragraph incorporating these ideas. For completeness, here is the amended text followed by the new text from this section (lines 252-268):

> "Since NSE puts more weight on large flow (Criss and Winston, 2008), it is not surprising that our NSE-log10 scores are better because extremely wet months are weighted less than using the raw runoff values compared to raw NSE scores. It is also notable that the VE scores using log10 values across the entire dataset is are better (0.363 vs. 0.733; Table 2). The physical significance of these observations could be that for large basins the time of concentration of any given flood event will be much longer, thus flood peaks will be wider and subdued due to infiltration into the shallow aquifers. However, a more likely explanation is that the average rainfall intensity is too low in the GSWP3 precipitation data used in producing GRUN. A bias at high and moderate flow conditions is particularly evident in the flow duration curves (Figure 5) and is likely due to downscaling from averages over larger areas with topographic complexity. While GSWP3 uses downscaled 20th century reanalysis products (Kim et al., 2017) at T248 resolution (~0.5º, the same resolution as GRUN), the topographic complexity of tropical islands such as the Philippines on the sub-0.5º scale would likely results in smoothing of the variability and lowering of the absolute magnitude of the precipitation fields, particularly during the wet season and during large synoptic precipitation events during the monsoon season. 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 because of averaging of the input products and/or discharge datasets to monthly values. Finally, while the GSWP3 precipitation inputs are bias-corrected using the Global Precipitation Climatology Centre precipitation network, previous work has highlighted data quality issues with some historical data from the Philippines (Schneider et al., 2014)."

Regarding the second place that we suggested this: we would like to keep the possibility and previous text that shallow aquifer filling could be important in certain locations, we have unpublished stable isotope data we are currently working on for publication (see Ibarra et al., 2020 Goldschmidt and AGU talks) suggesting fairly significant 'old' water in small and medium size catchments from the Luzon area (following methods from papers by Jasechko and Kirchner). We did however add a reference to the above at the end of this paragraph noting our previous description of the potential problem with the rainfall forcing for this region in GSWP3/GRUN (line 296-297):

> "Alternatively, this could be due to an underestimation in the wet season rainfall products from GSWP3 used to create GRUN as described in the previous section."

L287: I thank the authors for taking up the idea of bootstrapping but think that it is not done correctly here. Taking out individual months from a range of catchments is likely not so helpful because of the large amount of 'redundant data' in long time series. The question is how sensitive the bias correction factor is to the choice of the catchments or the number of catchments for which data are available. Thus instead of randomly taking out data points (from different times and different catchments), it would be better to exclude all the data from a certain number of catchments and to then determine how this affects the bias correction factor and the uncertainty in the bias correction factor. In fact, I would suggest that the authors do not only take out a fraction of the catchments for the bootstrapping but also test what the uncertainty of this factor would be if they had only data for one (or two or three or five) catchments per climate zone. This would be helpful for readers from other countries who may not have access to data from so many catchments to determine a bias correction factor.

This was an excellent point that we have overlooked, thank you for the suggestion. The previous reviewer suggested the individual months be extracted. We re-did the analysis by removing based on catchment rather than random month to overcome the problem of redundant data. We reflected on this comment at length and carried out the analysis on a country-wide basis for 1 to 55 catchments, we show this for 1 to 25 in the new Figure 7, where ~20 is where the bias coefficient metrics asymptote to the mean values, which is the number of catchments used when we report our estimated uncertainty in the text. Further, using a leave-one out approach we assess our bias correction's uncertainty further in a new sentence. This new approach to the bias correction, which does not change the mean values of the coefficients, but helps to quantify the uncertainties and the limitations shows that the mean values of the bias correction coefficients do not asymptote until at least 10+ catchments, an important finding that we also highlight in a new sentence.

Thank you for this suggestion, we hope we have implemented this idea in a satisfactory manner. Based on the above we have rewritten this section (lines 298-312):

> "Given the biases and in particular the clear underprediction of streamflow in GRUN during the wettest months we perform a bias correction of the GRUN dataset at a nationwide level using all the available data used in our analysis. We do so in a two-step process to both correct the mean offset and stretch the wettest months to higher values with all transformations occurring in log-transform space (i.e., as displayed in cross plots in Figures 2 and 4). Thus, we first add the mean

log10(runoff) difference between the observations and the predicted values (0.117 ± 0.045). Following this, using the *lm* function in R, we fit a linear regression between the observations and the GRUN predicted values (log10(runoff, observed) = m × log10(runoff, predicted) + b) and correct the predicted values using the slope (m=0.774 ± 0.058) and intercept (b=0.099 ± 0.030) derived from this regression. Uncertainties reported here are 68% confidence intervals and were assessed by bootstrap resampling observation and prediction pairs from 20 catchments (vertical line in Figure 7) without replacement 10,000 times. In Figure 7 we show the influence of including an increasing number of catchments from the dataset in our bootstrap resampling to assess how the mean value of the coefficients asymptote as more catchments are included. While the mean log10(runoff) offset is relatively unaffected, the slope and intercept of the bias correction do not asymptote until more than 10 catchments are included in the analysis. Finally, a leave-one-out approach (i.e., calculating the coefficients with 54 of 55 catchments) indicates 68% confidence uncertainties of ± 0.005, ± 0.006, and ± 0.003 for the log10(runoff) offset, the slope and intercept, lower than those reported above, as expected."

L291-293: This requires some rewriting as the text and the logic are difficult to follow.

Agreed. We have simplified and rewritten this sentence by breaking it up into several sentences, this was again a suggestion of the other reviewer:

"Because these corrections were carried out in log10 space statistical bias in the form of underestimation is possible (Ferguson, 1986). Following Ferguson (1986) we calculate the unbiased estimate of the variance (notated as 's') as 0.0686 mm/day which gives a correction factor (calculated as $\exp(2.65s^2)$ of 1.0126. This correction factor, a multiplier, can be applied to the bias corrected values to adjust for possible the bias due to the log10 space regression we have implemented."

L324: This sentence is not clear. Are you really suggesting that even though the GRUN database was not intended to be used for predicting flow for individual catchments, it can be used that way after bias correction? I don't think that you can conclude this based on your results!!

Sorry for the confusion. We agree that based on our analysis it cannot be used for individual catchments. We have clarified this based on the reviewer's suggestions and rephrasing. The last two sentences now read:

"GRUN was not intended to be used for estimating discharge for single small catchments, however it is applicable for use in regional and country scale analyses provided that proper statistical comparison of modelled versus actual gauged data are performed. We thus propose that the use of the GRUN dataset can be extended to other ungauged tropical regions with smaller catchments after at least applying a similar correction as described in this study."