Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-26-AC2, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Technical Note: Evaluation and bias correction of an observations-based global runoff dataset using historical streamflow observations from small tropical catchments in the Philippines" *by* Daniel E. Ibarra et al.

Daniel E. Ibarra et al.

dibarra@berkeley.edu

Received and published: 5 July 2020

Response to Anonymous Referee #2 by Ibarra, David and Tolentino [Note our responses are listed as "RESPONSE: ...."]

Ibarra et al. compare the monthly streamflow data from a new global database (GRUN) with gauge data for multiple catchments in the Philippines. The catchments were not included in the development of the global database and are smaller than those used

to train the machine-learning algorithm to create the global database. The work is very interesting because it highlights issues with the use of such global databases for smaller catchments or to answer local questions.

I thank and congratulate the authors for putting together a very valuable database of streamflow for catchments in the Philippines. This is highly useful because so much of our collective research efforts and knowledge focuses on catchments in temperature climates.

Nonetheless, I do have several comments (see below). These mainly focus on the lack of a comparison with a lower (and upper) benchmark, the use of overall statistics and the effects of pooling of data with different record lengths and variability on these overall statistics, the over-use of log-log transformation and scales, the overall message, and the appropriateness of the technical note category.

RESPONSE: Thank you for the encouraging, constructive and thorough review. We respond below to each of these points in detail.

I include other detailed comments and suggestions to strengthen the manuscript in the attached pdf. This pdf also has some editorial suggestions. Note that these are just suggestions to make the text more to the point or clearer - I don't think that all of them need to be addressed or implemented.

RESPONSE: We thank the reviewer for edits in the attached pdf, all of these editorial suggestions will be taken/addressed in our revision.

1. To assess the value and skill of the GRUN database, the results need to be compared to a lower benchmark. How well (or poor) does a very simple estimate reflect the observed monthly streamflow and how much better are the GRUN estimates than this rough estimate or lower benchmark? This lower benchmark could be based on the multiplication of the average runoff ratio for the region and month by the local rainfall data or the average streamflow from all catchments in a region. Similarly, we can not

expect a perfect r2 or NSE score because the observations are uncertain. Although the uncertainties are mentioned for the data in one database, they are not shown or discussed anywhere. As a result it remains difficult to interpret the results, i.e., should we consider these r2, VE, NSE or RMSE values as poor or does the GRUN data have some reasonable predictive power that is much better than a regional average value?

RESPONSE: Regarding the lower benchmark, we view the NSE as a good estimator of a lower benchmark. Given that the NSE value before bias correction is quite poor (country-wide) with a value of 0.091 (and a log10 value of 0.453), and an improvement in this score following bias correction, we view this as significantly better than the lower benchmark of just using the mean value of the observed data. Unfortunately, we have a limited number of paired rain gauge and catchment discharge data to perform a runoff ratio calculation as suggested by the reviewer. However, the use of satellite-based rainfall estimates calibrated with limited rain gauge data coupled to GRUN discharge data is the subject of another paper our group is currently preparing.

RESPONSE: Regarding the metrics used: as a point of comparison the r2 and NSE values that are calculated and reported in our Table 2, are lower than again the large GRDC river basins compared to in the original GRUN paper (Ghiggi et al., 2019; see their Table 2). This is an important point of our paper, even once bias corrected using the compiled datasets the NSE value is lower than that of the global comparison. This is important as it demonstrates the need to incorporate smaller tropical catchments such as those presented here in such global runoff datasets, particularly with respect to correctly predicting high-flow during wet seasons. To address this point, we will add further discussion to the paragraphs starting in line 274 with a direct comparison to the original GRUN paper.

2. Although the time series of observed and GRUN based streamflow are given in the supplementary material, none of these time series are shown in the main document. I highly recommend showing these plots for the best and the poorest site in each climate zone in the main document. This will give the reader a much better feeling of the skill

of the GRUN data and helps with the interpretation of the VE, r2 and RMSE values that are given in the text.

RESPONSE: To address this comment we will add a new Figure (between figures 1 and 2 currently) showing 8x examples from the existing supplementary figures (time series and cross plots), listing the VE, r2 and RMSE values for that catchment.

3. A large part of the analyses is based on pooled data (e.g. Figure 3) but the record length is very different for the different catchments and the variability in discharge is also different. This likely influences these pooled results. I would rather (also) see boxplots that show the r2, NSE and NSE-log values for the individual catchments as is done for VE in Figure 4. This will also give the reader a much better feeling of how different these results are for the different catchments. I therefore suggest to add these plots to the manuscript and to add these ranges to Table 2 as well.

RESPONSE: We agree to the comment thus we placed importance in including individual metrics so readers can see and reinterpret biases in the dataset as what reviewer 2 has done. We will include similar box plots of r2, NSE and NSE-log values for the individual catchments as in Figure 4 for VE and add additional lines with the median and IQR to Table 2. We append a draft of this figure to this reply, this will be added to Figure 4.

4. Almost all comparisons of the observed and GRUN-based discharge are shown in log-log space. This is informative for some analysis and allows one to see the data but at the same time almost any comparison looks OK in log-log space, even when the data don't really match. I therefore suggest to not use log-log axes where it is not entirely necessary. For example figure 1a could be split in 3 sub-panels (max, median,min) and then show the data on a linear scale. Furthermore, I wonder whether the bias correction in log-space leads to large errors when the corrected values are transferred again. This isn't shown nor discussed in the manuscript.

RESPONSE: This is a good comment that was also brought up by the first reviewer.

To rectify we will: - Split Figure 1a into 3 sub-panels in linear scale for the max, median and minimum runoff. - Bias correction via log-space does introduce errors, particularly for large runoff values with ratio of the true to predicted values scaling approximately as exp(2.65sˆ2) where s is the unbiased estimator of the variance (see Ferguson (1986) referred to us by the first reviewer). To account for this we will both perform a bootstrap analysis (shown above) and then mention in the text that it is suggested to also apply this scaling and provide the value for the unbiased estimate of the variance (s). Ferguson, R. I. River Loads Underestimated by Rating Curves. Water Resour. Res. 22, 74–76 (1986).

5. I think that the overall conclusion of the manuscript is too optimistic (although this admittedly depends on the comparison with the lower benchmark (see comment 1)).I agree that the bias correction helps and leads to a significant improvement but the GRUN based estimates of streamflow are very poor (particularly when one looks at the time series that are given in the supplementary material). The abstract and conclusion could highlight the danger in using these types of products for local streamflow prediction, model calibration, etc more clearly. Now it seems to be overly optimistic on how these data can be used for a range of local studies or to answer local questions.C3

RESPONSE: We agree that our overall conclusions are overly optimistic, we will add the following to the conclusions: “Global databases such as GRUN are applicable for aggregated stream discharge estimates and analyzing for general trends in the hydrologic characteristics of a region. The recommended bias correction presented here will likely improve such estimates and analysis for the Philippines. While GRUN, a global runoff discharge dataset, was never intended to be used for estimating single catchment discharge and was not trained against smaller catchments as compiled here, its applicability for such purpose can be extended provided that proper statistical comparison of modeled versus actual gauged data are initially performed.

6. The manuscript was submitted as a technical note but the manuscript doesn’t fully fit the description of a technical note as it doesn’t describe a new method or technique.

It should thus be a regular research paper, which would allow for more comparisons of the datasets (as described above) and additional figures (as described above).

RESPONSE: We will consult with the HESS editor but do feel that this does fit the description of a technical note because we have compiled a new runoff dataset from a previously unrepresented country for the community to use and provided a tool via our bias correction equation for correcting a global observation-based gridded runoff dataset for use based on this data. That said, given the above suggestions we are happy to turn this into a regular research paper if the editor feels this is necessary.

7. Some more background information on the gauging station data used in GRUN would be helpful, e.g. what percentage of stations were located in the tropics? And do the papers that describe the GRUN database make claims about smaller catchments or tropical catchments?

RESPONSE: We agree with the reviewer that these are important details regarding GRUN that we overlooked mentioning, we will add the following details to the revised manuscript: - GRUN is highly biased to the northern hemisphere mid-latitudes. We will refer the reader to the original publication (Ghiggi et al., 2019) with respect to details mentioning the underrepresentation of small tropical catchments (not just due to the criteria of >10,000 km2 but also the lack of records in general). - Additionally, we will mention that the GRUN publication does discuss that because of how they trained the dataset the uncertainty is greatest the tropics (see bottom left of page 1164 and Figure 8b of Ghiggi et al., 2019) in comparison to other datasets, and mention southeast Asia showing an increase in runoff rates over the period of analysis despite a paucity of records (see top right of page 1666). - Further, a strong correlation of runoff with ENSO is shown in figure 11 of Ghiggi et al. (2019) in southeast Asia, though the resolution of the map makes it difficult to ascertain significance and regional variability in the Philippines. We thank the reviewer for these questions about GRUN, while not our group’s product it will certain strengthen our manuscript.

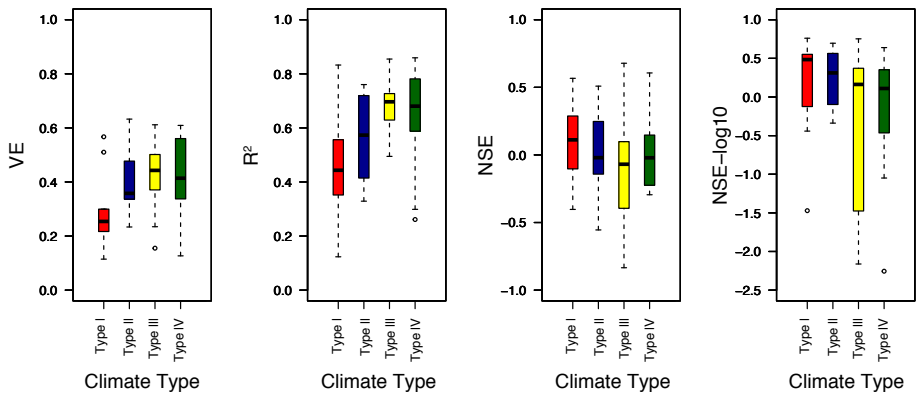

Fig. 1. Draft of additions to Figure 4 showing VE (previously panel 4b) as well as additional scores (R2, NSE and NSE-log10).