Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-26-RC2, 2020

Hydrology and Earth System Sciences
Open Access
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.

Anonymous Referee #2

Received and published: 7 June 2020

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.

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.

Major comments:

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?

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.

3. A large part of the analyses is based on pooled data (e.g. Figure 3) but the record length are 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.

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.

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.

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).

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?

Please also note the supplement to this comment:
https://www.hydrol-earth-syst-sci-discuss.net/hess-2020-26/hess-2020-26-RC2-supplement.pdf

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