Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2017-508-RC2, 2017

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on “Global-scale evaluation of 23 precipitation datasets using gauge observations and hydrological modeling” *by* Hylke E. Beck et al.

Anonymous Referee #2

Received and published: 12 October 2017

General comments:

Beck et al. evaluated the performance of 23 precipitation datasets using gauge observations and a HBV hydrological model. The paper fits very well within the stated scope of journal and I read the paper with great interest. The authors deserve considerable credit in taking this extensive study and producing a concise manuscript.

However, I would like to address some suggestions:

- I believe that this manuscript will become more useful if the authors can give further breakdown and more deep analyses for their result in Table 2, e.g. by classifying it to several continents/regions or several climate regions (e.g. as done for Table 3).

- The authors used only NSE for their evaluation using HBV model calibration (while they used several metrics for evaluating P datasets to gauge observations). I am just wondering why the authors selected NSE (among many other measures) for their calibration exercise.

One of the concerns of using NSE, which is a normalization form of the mean squared error (MSE), is its reputation that emphasizes high flows (Legates and McCabe, 1999; Krause et al., 2005). The disadvantage of NSE is the fact that errors between observed and modeled values are calculates as squared values. Consequently, NSE is overly sensitive to large values in time series (whereas lower values are less important). Gupta et al. (1999) mentioned other weaknesses of NSE. One of them is the fact the bias component in NSE is normalized by the standard deviation (i.e. variability) in the observed flows. This means that the bias in time series with high flow variability tends to have little influence in the optimization of NSE, possibly leading to simulations having large volume balance errors. There are many other studies (see e.g. Schaefli and Gupta, 2007) discussing potential problems of using NSE and even Beck et al. (2016) acknowledged NSE as a week metric.

Note that by providing this comment, I am not necessarily suggesting that the authors have to repeat their calibration exercise with different objective functions (which may be very computationally expensive). Rather, I would like to recommend that the authors should validate their existing calibrated runs (already chosen based on their NSE optimization) by calculating some other metrics, e.g. KGE, MAE (mean absolute error), or log NSE. I believe that such validation will make this study more convincing. One can even speculate that an evaluation using log NSE, which emphasize low flow periods, may confirm one of their findings: the superiority of the MSWEP datasets v2.0, which has the best performance in terms of annual dry day error (Table 2).

Details / specific comments:

Section 2.1: I suggest that the authors add brief description for each P dataset. I believe that this will help readers and improve the quality of the manuscript. Such an explanation can be relatively short as there are similar datasets that can be grouped together, e.g. CHRP and CHRPS, CMORPH and CMORPH-CRT, and all MSWEP datasets.

Table 1: - Please also clarify what the difference between Land and Global. Does the latter include ocean? - Please also explain in the text about the subscript –ng for MSWEP.

Section 2.2: - Page 5, lines 5-7. Here you decided to use MAE, instead of RMSE. I am just wondering why you used NSE, a similar criteria as RMSE, for your performance evaluation using hydrological modelling (Section 2.3)?

Section 2.3: - Why did you use NSE? - Why did you use exclude large catchments (> 50,000 km2)? - If there are several stations along a river (e.g. Meuse), did you use only the most downstream one? Please clarify.

Table 2: Further breakdowns into several continents or climate regions will be useful.

Section 3.1: - Page 7, lines 1-2: MSWEP V2.0 obtained substantially lower mean annual P trend errors than the other P datasets (Table 2 and Supplementary information Figure S5). Please remove “substantially” as the range of these errors is relatively small (as also stated in lines 11-12). - Related to annual P trend errors, I am also wondering what the results will be if longer time series (e.g. starting from 1981) are used.

Section 3.2: I believe that it is more useful to classify and analyze the performances over different climate regions (or continents).

Page 7, line 29: I am curious with the paper Beck et al. (2017a), which is still in preparation.

Section 3.3: - Page 8, lines 23-26. This just shows the superiority of MSWEP datasets. Can you please confirm this superiority for other metrics, e.g. KGE and log NSE. - I am

also wondering what the results will be if longer time series (e.g. starting from 1981) are used. Can you please discuss this?

Table 3: Please improve the caption. What do the letters A, B, C, D and E stand for?

References:

Beck, H. E., A. I. J. M. van Dijk, A. de Roo, D. G. Miralles, T. R. McVicar, J. Schellekens, and L. A. Bruijnzeel, 2016, Global-scale regionalization of hydrologic model parameters, Water Resour. Res.,52, 3599–3622, doi:10.1002/2015WR018247.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009, Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91.

Krause et al., 2005, Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5(89), 89–97.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model evaluation. Water Resources Research 35, 233–241.

Schaefli, B., and H. V. Gupta, 2007, Do Nash values have value?, Hydrol. Processes,21(15), 2075–2080

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-508, 2017.