Dear Authors,
Ref#1, whom I warmly thank, has carefully reviewed the revised manuscript. The Referee has greatly appreciated your revision work and suggests a final list of minor amendments that will further improve its presentation to our readership. I am therefore inviting you to address such final, constructive comments and look forward to reading the last version of your manuscript (that I will ask also Ref#1 to review for one more time, if they may find the time).
Best wishes,
Elena Tothj
University of Bologna
HESS Topical Editor

Response : Thank you for providing us with the opportunity to further revise and improve our manuscript. We have revised the manuscript in accordance with the referee's comments. The major revisions include an expanded discussion on the low performance of HBV in arid regions and the addition of new sections in the Supplementary Material. Please find our detailed responses to each comment below.

Minor comments:

a) Previous MC5. Code availability. The authors stated that "We have added the documentation on how to use the code and run it with user-defined input files, with examples within the examples folder." However, the file available at https://github.com/AtrCheema/rain2flow/blob/master/examples/readme.txt is currently empty. Please add at least minimal content to this file to facilitate the use of the Python code together with the examples provided in the `examples` folder.

Response : We added the example code to run hbv for a test catchment and the online run of the example is available at https://rain2flow.readthedocs.io/en/latest/auto_examples/hbv.html . We have explicitly added the link to this notebook in that readme.txt file. Furthermore, the example code is also provided on the readme file of GitHub repository.

b) Previous MC10. Selection of temporal period used for the calibration of the individual catchments. I thank the authors for their reply to my request to use the same temporal period for the calibration of all precipitation (P) products in each catchment, as well as for the statement added to the Potential Limitations and Future Work subsection. However, to ensure that readers are fully aware that the KGE values obtained for each product within a given catchment are not directly comparable (due to the use of different calibration periods), I request that the authors add a clarifying note to the legend of Figure 2 indicating that the calibration periods used for different datasets may differ in length. I suggest changing the current text:

“Calibration KGE, correlation (r), long-term bias (β), and variability ratio (γ) scores achieved by the 24 P datasets.”

to:

“Calibration KGE, correlation (r), long-term bias (β), and variability ratio (γ) scores achieved by the 24 P datasets (for each catchment, the calibration period for each P dataset was not necessarily the same).”

Response : Thank you for your suggestion. We have modified the caption of Fig. 2 by adding “For a given catchment, calibration periods were not necessarily consistent across P datasets because their temporal coverage differs.”

c) Previous MC11. I thank the authors for their reply regarding the unexpected result that the CPC Unified dataset, which is based solely on rain gauge observations, ranked third among all datasets (based on KGE only). However, I request that the figure used in their reply (unnumbered, with caption “Comparison of P dataset performance for catchments with high rain gauge densities vs catchments with low rain gauge densities”) be included in the supplementary material and referenced in the revised manuscript (Lines 218–220).

Response : Thank you for your suggestion. We have added this figure in supplementary (Fig. S29) material and referenced it in the revised manuscript.

d) Previous MC12. I thank the authors for their reply regarding the ability of the P products to represent mean annual precipitation. However, considering that in the lower-right boxplot of Figure 2 (Bias, β), the lower whisker for GSMaP V8 is nearly as extended as that of IMERG-Early V7, I request that GSMaP V8 be explicitly included in the revised text (Lines 263–266), as follows: “We found that several satellite P datasets (notably IMERG-Early and -Late V7, GSMaP V8, SM2RAIN-ASCAT, SM2RAIN-CCI, and CMORPH-CDR) exhibit …”

Response : Thank you for your suggestion. We have modified these lines as suggested.

e) Previous MC15a. Poor performance of HBV in arid climates. I appreciate the addition of the new text in Lines 282–288 and 273–285, particularly the statement:

“This lower performance does not necessarily reflect an inability of HBV to represent arid hydrology; rather, it reflects a general decline in skill across hydrological and land surface models in such regions (e.g., Beck et al., 2017a).”

While I generally agree with this statement, it is important to also acknowledge that “more effort should be devoted to calibrating and regionalizing the parameters of macro-scale models” (Beck et al., 2017a), and that "calibration of TUWmodel (an HBV-like model) was able to compensate, to some extent, for differences in annual and intra-annual P amounts, intermittency, and extremes (see Figs. 2 and 3) among the four products" (Baez-Villanueva et al., 2021).

Response : Thank you for your suggestion. However, even after calibration, the performance is poor in arid regions, as demonstrated in our study as well as other studies (e.g., Beck et al., 2017a, 2016). The statement that “more effort should be devoted to calibrating and regionalizing the parameters of macro-scale models” refers to the fact that many macro-scale models are uncalibrated.

After careful consideration, we have revised this part as follows:

Lines 285 - 291: *“In arid climates, all P datasets tend to perform relatively poorly, with a slight advantage for (re)analyses over satellite-based datasets, consistent with previous evaluation (e.g., Beck et al., 2017a, 2016). The lower arid-region scores mainly reflect (i) poorer forcing quality due to the short-lived, localized nature of storms and sub-cloud evaporation (virga) (Wang et al., 2018); (ii) more threshold-driven runoff generation that amplifies small forcing errors; and (iii) fewer runoff-producing events, which increases sampling uncertainty (Beck et al., 2017a, c; Sun et al., 2018; El Kenawy et al., 2019; Beck et al., 2019a; Williams, 2025). Thus, the lower performance does not necessarily indicate an inability of HBV to represent arid hydrology.”*

References :

Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*(5), 3599-3622.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, Hydrology and Earth System Sciences, 21, 2881–2903, 2017a.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F., and Wood, E. F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS, Hydrology and Earth System Sciences, 23,207–224, 2019a.

El Kenawy, A. M., McCabe, M. F., Lopez-Moreno, J. I., Hathal, Y., Robaa, S. M., Al Budeiri, A. L., Jadoon, K. Z., Abouelmagd, A., Eddenjal, A., Domínguez-Castro, F., Trigo, R. M., and

Vicente-Serrano, S. M.: Spatial assessment of the performance of multiple high-resolution satellite-based precipitation data sets over the Middle East, International Journal of Climatology, 39, 2522–2543, https://doi.org/10.1002/joc.5968, 2019.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L.: A review of global precipitation datasets: data sources, estimation, and intercomparisons, Reviews of Geophysics, 56, 79–107, 2018.

Wang, Y., You, Y., and Kulie, M.: Global Virga Precipitation Distribution Derived From Three Spaceborne Radars and Its Contribution to the False Radiometer Precipitation Detection, Geophysical Research Letters, 45, 4446–4455, https://doi.org/10.1029/2018GL077891, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL077891, 2018.

Williams, G. P.: Friends don't let friends use Nash-Sutcliffe Efficiency (NSE) or KGE for hydrologic model accuracy evaluation: A rant with data and suggestions for better practice, Environmental Modelling Software, 194, 106 665, https://doi.org/https://doi.org/10.1016/j.envsoft.2025.106665, 2025.

f) Therefore, I request that the authors include additional discussion regarding the role of model calibration in the lower performance of HBV in arid regions . The revised text could be something similar to:

"This lower performance does not necessarily reflect an inability of HBV to represent arid hydrology; rather, it reflects generally low skill across land surface models in such regions (e.g., Beck et al., 2017a), which could be partially compensated through catchment-specific calibration of the hydrological model (Baez-Villanueva et al., 2021)."

Reference

• Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Mendoza, P. A., McNamara, I., Beck, H. E., Thurner, J., Nauditt, A., Ribbe, L., and Thinh, N. X. (2021). On the selection of precipitation products for the regionalisation of hydrological model parameters. *Hydrology and Earth System Sciences*, 25, 5805–5837. https://doi.org/10.5194/hess-25-5805-2021.

Response : As requested, we have added reasons for the lower performance in arid regions. Please see our response above. Calibration may indeed partially mitigate the lower performance

in arid regions, but we did not explicitly compare uncalibrated and calibrated scores. Therefore, we cannot conclude that calibration provides greater compensation in arid regions than in more humid regions. We believe the revised explanation above clearly summarizes the key factors reducing performance in arid regions.

g) New minor comment. Considering that the supplementary material has 56 figures numbered from S1 to S56 (with some new figures expected after this new review round), I ask the authors to provide some structure to this supplementary material (e.g., adding some sections while keeping the figure numbers), in order to make it more useful for the reader.

Response : We have added sections to the supplementary material to improve readability. Thank you for the suggestion.