Dear Authors,
as I expected, the Referees are both very satisfied with your revised manuscript, that they consider excellent both for the content and the presentation.
Referee #2 has listed a set of minor revisions that will further improve the quality of your presentation and its value for the journal readers, so I warmly invite you to make these last improvements and we both look forward to reading the last version of the manuscript.
Best wishes,
Elena Toth
University of Bologna
HESS Topical Editor

Response: Thank you for the opportunity to further revise and improve our manuscript. Please note that we have revised the captions for Fig. 3 and Table 3 and have also improved the grammar and clarity throughout the manuscript. All changes have been tracked and are provided in the "track-changes" file. Please find our responses to each reviewer comment below.

Report # 2

I thank the authors for having provided a thorough reply to the comment made by the three reviewers. In particular, I am glad to hear that you updated your streamflow dataset, fixed some bugs, changed the potential evapotranspiration formula from Hargreaves-Samani to Penman-Monteith (with additional data requirements), removed CMORPH-RAW and added CMORPH-CDR, adding four different scenarios for analysing the calibration of PCORR and SFCF parameters of the HBV model, among others. In my opinion this new version of the manuscript will be of greater help to the global hydrometeorological scientific community. However, some comments were not fully addressed in your reply, and therefore I now recommend a minor revision. The following lines describe the major and minor problems detected in the manuscript.

Response: Thank you very much for acknowledging the changes in the manuscript. We have revised the manuscript as per your comments. Please find the responses to each of your comments below.

Major comments:

1. Previous MC4. Catchment selection. To ensure the suitability of the catchments used in the analyses, five selection criteria were applied in the manuscript to the 34,768 streamflow stations that passed the duplication check. In their reply, the authors correctly replied to the comment related to discarding streamflow stations where both the station location and the corresponding catchment centroid were within 5 km of those of another station. However, they did not reply to the comment related to the adoption of a number of events (defined as runoff >5 mm d−1) larger than 10 non-consecutively, to ensure sufficient data for calibration. Could you provide one or two lines explaining this criterion?. In particular, what should be understood by "non-consecutively"?, what is the minimum duration of an event? and, why not using fewer but longer events to provide sufficient data for calibration?

Response: This was indeed not clearly explained. We have revised the text as follows to avoid confusion:

Lines 101 - 103: *"The number of days with appreciable runoff (>5 mm d^{-1}) had to exceed 10, and these days could not be consecutive (i.e., they should not be part of a single continuous event). This ensures that the calibration is based on a sufficient number of distinct runoff events."*

We hope this clarifies the criterion. Thank you for bringing this to our attention.

2. Previous MC5. I thank the authors for having taken the decision of open source the Python code of their HBV model. However, the link provided in their response (https://github.com/hyex-research/hydro_clim_scen_analysis/blob/main/hbv.py) currently point s to a "404 error " (page not found). In addition, the previous link does not coincide with the link provided in the "Code availability" section of the new version of the manuscript (https://github.com/AtrCheema/), which currently points to two test files (i.e., no to the actual model code) that do not allow to run the HBV model with user-defined input files. Therefore, I request to provide the link to the model code files with a bare-minimum documentation about how to run the HBV model with user-specific input files.

Response: The correct link to the HBV code which is used in this study is https://github.com/AtrCheema/rain2flow . 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. Please note that the link in the 'Code availability' section points to this correct link.

3. Previous MC10. Selection of temporal period used for the calibration of the individual catchments. In their reply, the authors mention that "we used he full period of overlapping streamflow and P data for each catchment. Using a common temporal range for all P datasets and all catchments would significantly decrease the number of P datasets and catchments considered in this study, and would result in less generalizable results". I agree with them that using a unique temporal range for all P datasets and all catchments would lead to an important decrease in the number of P datasets and catchments considered in this study, which is not desirable. However, my original request was "to use the same temporal period for the calibration of all P products in each catchment", i.e., for each catchment to use a unique temporal period to calibrate the HBV model with all the P products, but this period might be different from catchment to catchment. I think this is the only way to ensure a fair comparison of the results obtained in a catchment, as the data length used in the calibration step has an important influence in the final model performance. Therefore, I still request to use a unique temporal period to calibrate the HBV model with all the P products, but this period might be different from catchment to catchment.

Response: Thanks for the clarification. Unfortunately, we cannot *"use the same temporal period for the calibration of all P products in each catchment,"* because several datasets have much shorter record lengths (e.g., GDAS), and some have no temporal overlap with one another at all (e.g., GDAS and GPM+SM2RAIN). Harmonizing the calibration period across all datasets within each catchment would require excluding short-record datasets, removing many otherwise suitable catchments, and forcing unnecessarily short calibration periods for long-record datasets. Each of these steps would reduce the scope and value of the study. Furthermore,

given the large number of catchments in our analysis, the influence of varying calibration periods on averages is minimal. Moreover, many datasets share broadly similar temporal coverage, and for these products the KGE values are fully consistent.

To explicitly acknowledge this issue, we have added the following statement to the Potential Limitations and Future Work subsection:

Lines 397 - 402: *"Some P datasets (GDAS and CMORPH-RT) have relatively short record lengths (Table 1), which can yield less stable KGE scores and may slightly overestimate performance, particularly in arid regions where P events are infrequent. Their limited temporal coverage also prevented the use of a single, uniform calibration period across all datasets. As a result, part of the variation in calibration performance may reflect differences in calibration periods rather than dataset quality. Nevertheless, given the large number of catchments analysed, the impact on the aggregated results and main conclusions is expected to be small."*

4. Previous MC11. In their reply, the authors correctly addressed the point related to datasets with higher spatial resolution do not necessarily result in better performance for hydrological modelling. However, the did not reply to my comment about the surprising fact that the CPC Unified dataset, which is based solely on rain gauge information, ranked second among all datasets, even in mountainous areas, where the density of rain gauges is typically low. Could you please a short discussion about this surprising result?

Response: Since CPC Unified is purely based upon rain gauge data, it ranked second best overall or in areas where the rain gauge density is high, which generally correspond to regions with more catchments. For areas with lower rain gauge density, the ranking of CPC Unified drops significantly. The following figure compares the performance of the *P* datasets in regions with different rain gauge densities. The upper plot shows the performance of the *P* datasets for the 100 catchments with the highest rain gauge densities, while the lower plot displays the performance of the *P* datasets with the lowest rain gauge densities. The ranking of CPC Unified drops to eighth place in catchments with the lowest rain gauge densities.

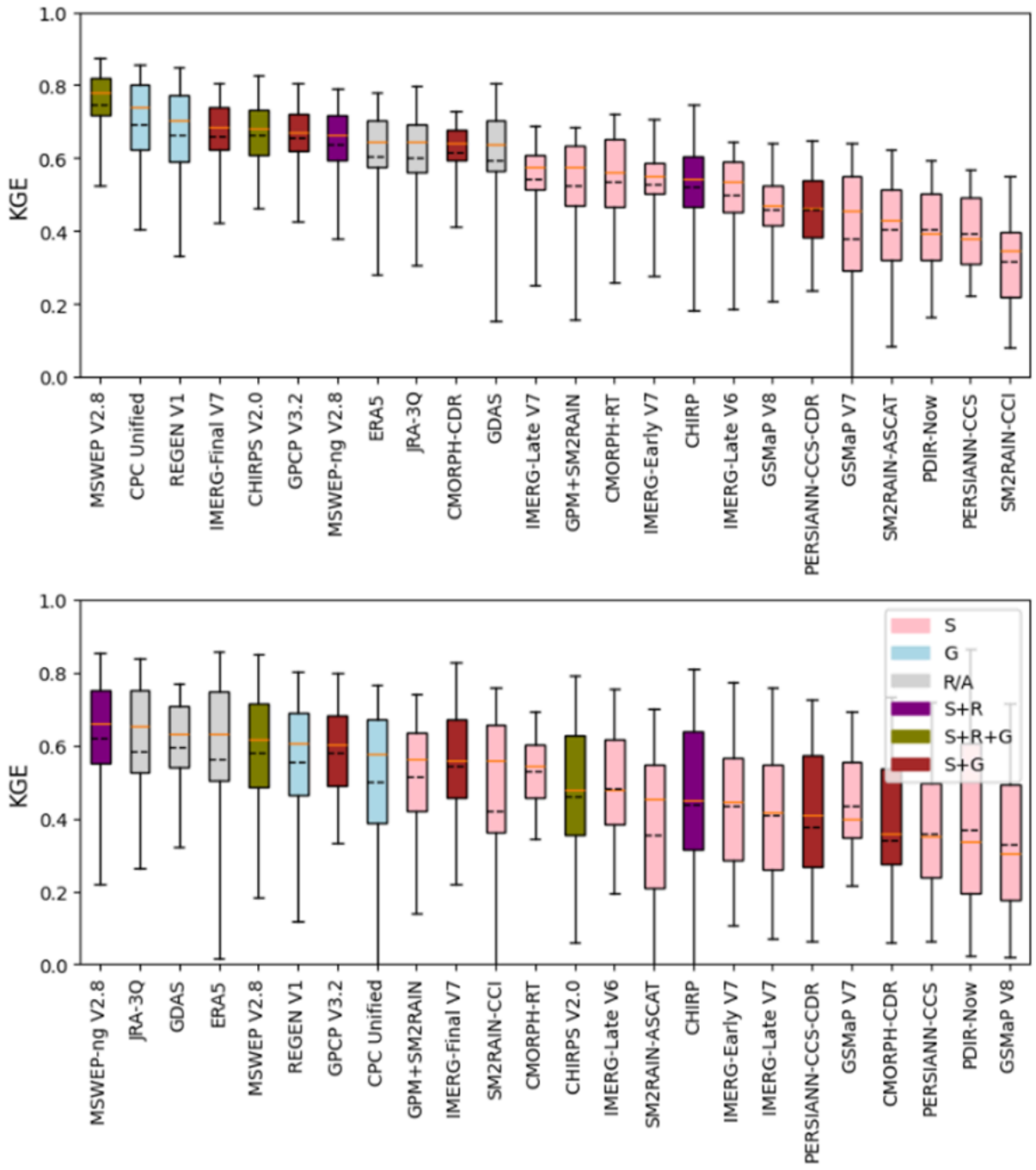

Fig. Comparison of *P* dataset performance for catchments with high rain gauge densities vs catchments with low rain gauge densities.

Related to CPC Unified performance in densely vs poorly observed regions, we have the following statement in the Overall Model Performance subsection:

Lines 218 - 220: *"Because CPC Unified and REGEN V1 rely exclusively on daily gauge*

observations, their performance is limited in these data-sparse regions, where values are interpolated between distant gauges."

And the following statement in the Limitations and Future Work subsection:

Lines 393 - 396: *"Since the global distribution of streamflow gauging stations closely aligns with that of meteorological monitoring networks (see Krabbenhoft et al., 2022, and Kidd et al., 2017), our assessment may slightly overestimate the relative performance of gauge-based P datasets and (re)analyses—which assimilate in situ observations from these networks—compared to satellite-only datasets.*"

5. Previous MC12. Based on the 21 figures provided in the author's reply, please provide a short discussion in the manuscript about the ability of some product to represent mean annual precipitation (e.g., PERSIANN-CCS, PDIR-Now, JRA-3Q, GSMaP V7), at least in comparison to the ensemble average value already computed.

Response: We appreciate the suggestion. The bias component of the KGE reflects the ability of the products to represent mean annual precipitation. We discuss the bias of several products throughout the manuscript, for example in lines 243 - 244. Furthermore, we have added the following lines linking the bias component of KGE and the 21 figures representing Pi - Pavg.

Lines 263 - 266 : "*We found that several satellite P datasets (notably IMERG-Early and -Late V7, SM2RAIN-ASCAT, SM2RAIN-CCI, and CMORPH-CDR) exhibit pronounced low-β tails (Fig. 2), indicating significant local P underestimation. This finding is further corroborated by maps of the difference between the mean annual P of each product and the multi-product mean (Supplement Figs. S32–S54), revealing extensive regions with negative values.*"

6. Previous MC14. I thank the figure included in your reply (caption: "Spatial maps and distribution of Kling-Gupta Efficiency for MSWEP V2.8 dataset").However, I think this figure would be useful to the wider scientific community to complement current figure 3. Therefore, I request to the supplementary material, with a proper link from the main body of the manuscript.

Response: Thank you for your suggestion. We added this figure to the supplementary material and referenced it in the main manuscript.

Lines 184 - 185 : "A more detailed spatial KGE map of the MSWEP V2.8 dataset in each catchment is illustrated in Supplement Fig. S1."

7. Previous MC15a. Could you please add some part of your reply about the poor performance of HBV in arid climates to the main body of the manuscript?

Response: Thank you for the comment. We have added the following text in the main body regarding the performance in arid climates:

Lines 282 - 288 : *"in arid climates, all P datasets tend to perform relatively poorly, with a slight advantage for (re)analyses over satellite-based datasets. P in arid regions tends to be brief and*

intense, making it challenging to detect and simulate accurately (Beck et al., 2017c; Sun et al., 2018; El Kenawy et al., 2019; Beck et al., 2019a). The occurrence of virga, or P that evaporates before reaching the ground, further complicates accurate P estimation in these regions (Wang et al., 2018).”

To emphasize that HBV is not to blame, we added the following sentence:

Lines 273 - 285: “. 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).”