This article presents a comprehensive evaluation of the daily performance of 23 global precipitation (P) datasets through a hydrological modeling experiment using the HBV model across 16,295 catchments worldwide. The manuscript is very well-written and clear, and the study addresses a topic of interest to the scientific community. It contributes valuable insights into the suitability of different P datasets for hydrological applications at the global scale and fits well within the scope of the journal. Overall, this is a strong and useful contribution, and I congratulate the authors for the scale and depth of the analysis. I believe that considering the following points will further enhance the manuscript.

Response: Thank you very much for your time to review the manuscript. We have revised the manuscript in accordance with your suggestions. We would also like to mention that we have used a revised and updated streamflow database, which includes a higher number of data sources (increased from 22 to 29), improved temporal coverage, and bug fixes—particularly for streamflow records from Africa (ADHI). The extended temporal coverage has led to an increase in the number of catchments with sufficient data for model calibration. As a result, the total number of catchments for which HBV parameters were optimized against observed streamflow increased from 16,295 to 18,428. We have also removed CMORPH-RAW and included CMORPH-CDR, following the recommendation of Reviewer 1. While the overall findings remain consistent with those in the previous version, we have updated the manuscript wherever new insights emerged. Please note that the line numbers and figure numbers in our responses correspond to the updated manuscript

I understand the rationale behind calibrating both the snowfall gauge undercatch correction factor (SFCF) and the multiplicative bias correction factor (PCORR), as these systematic errors can be addressed more easily. However, I suggest lowering the minimum bound of these parameters (e.g., to 0.6) to avoid favouring datasets that tend to underestimate P. In addition, it would be helpful to compare the results of the current calibration approach with a scenario where SFCF and PCORR are both fixed at 1.0. This comparison could shed light on the overall performance of P datasets. More specifically about (i) which datasets tend to systematically over- or underestimate and where, and (ii) the relative importance of these biases when these products are used for hydrological modelling purposes.

Response: We evaluated the performance of the precipitation datasets by calibrating KGE values under four different scenarios. In the first two scenarios, PCORR was allowed to vary between 0.0–2.0 and 0.5–2.0, respectively. In the third scenario, both PCORR and SFCF were fixed at 1.0, as suggested by the reviewer. The fourth scenario reflects our original setup, where both PCORR and SFCF were allowed to vary between 1.0 and 2.0. The resulting rankings of the precipitation datasets based on KGE values are illustrated in the figure below.

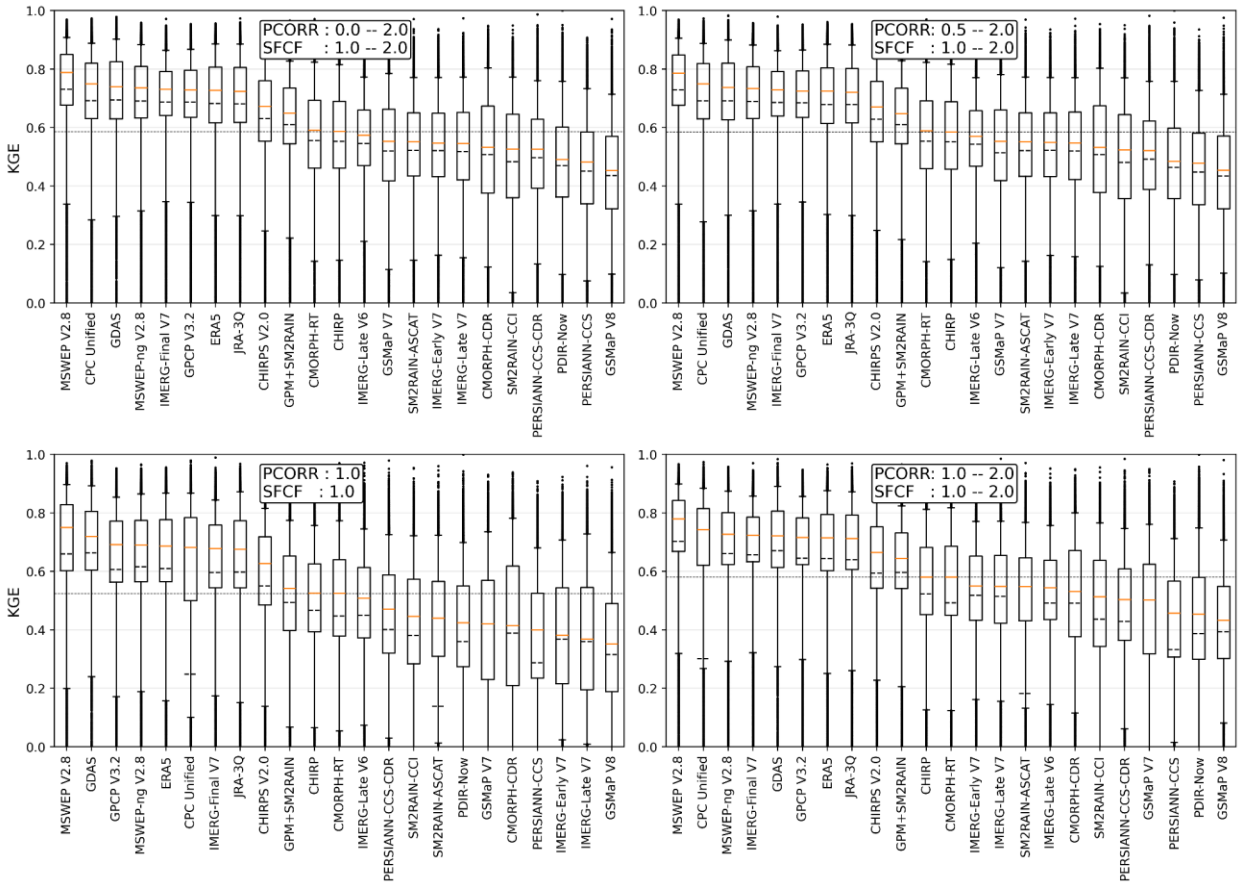

Fig S26. Kling-Gupta Efficiency values for P datasets in different scenarios of PCORR and SFCF calibration.

These results indicate that the overall ranking of precipitation datasets does not vary significantly across the four scenarios. However, in scenario three (PCORR: 1.0, SFCF: 1.0), the rankings of GPCP V3.2 and ERA5 improved. Notably, this improvement in ranking is not due to higher KGE values, as the overall KGE in this scenario is lower compared to the other three. The figure below illustrates the decrease in KGE when PCORR and SFCF were fixed during calibration (scenario three), compared to the default scenario where these parameters were allowed to vary between 1.0 and 2.0.

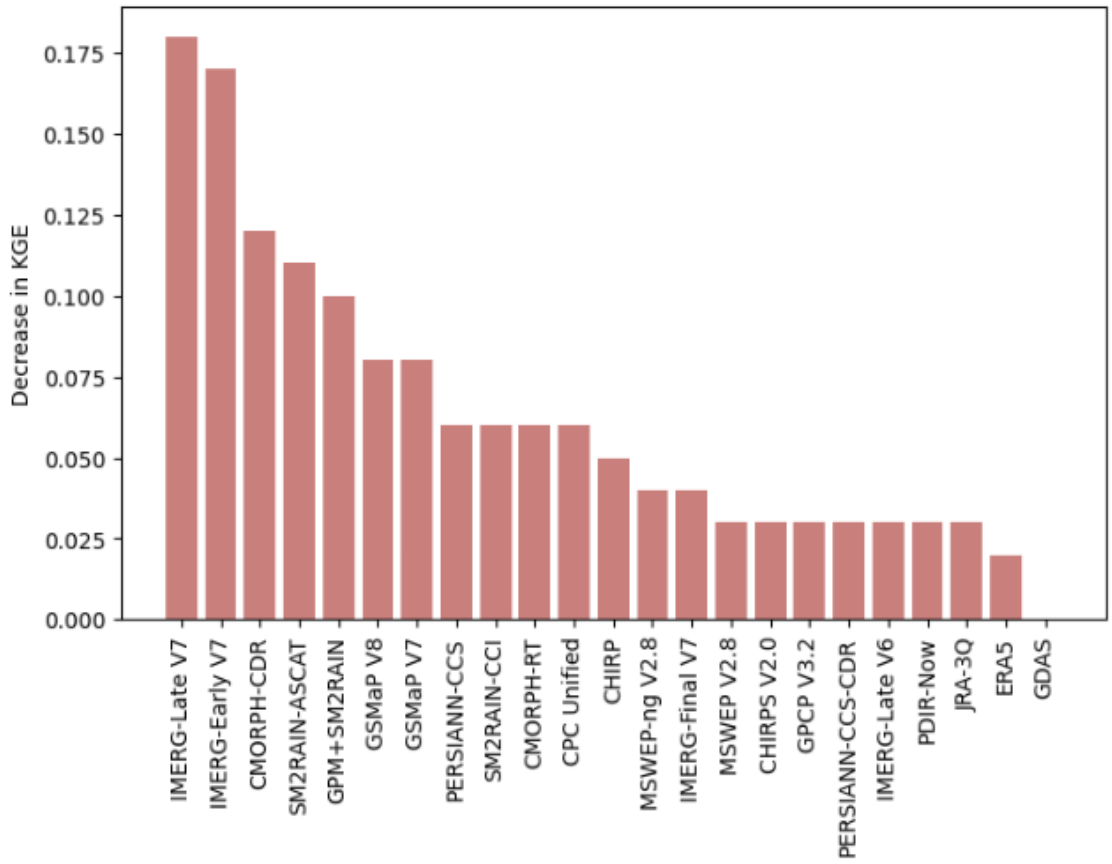

Fig. Decrease in KGE values when PCORR and SFCF were kept constant during calibration (1.0) as compared to when they were allowed to vary between 1.0 to 1.0.

The distribution of PCORR values across the four calibration scenarios (as shown in the following figure) reveals that only a few precipitation datasets exhibit significant adjustment below 1.0 when PCORR is allowed to vary from 0.0 to 2.0. Specifically, PDIR-Now, GSMaP V7, PERSIANN-CCS-CDR, and IMERG-Late V6 show a notable portion of calibrated PCORR values falling below 1.0. This indicates that, for these datasets, the raw precipitation input tends to overestimate actual precipitation in a substantial number of catchments.

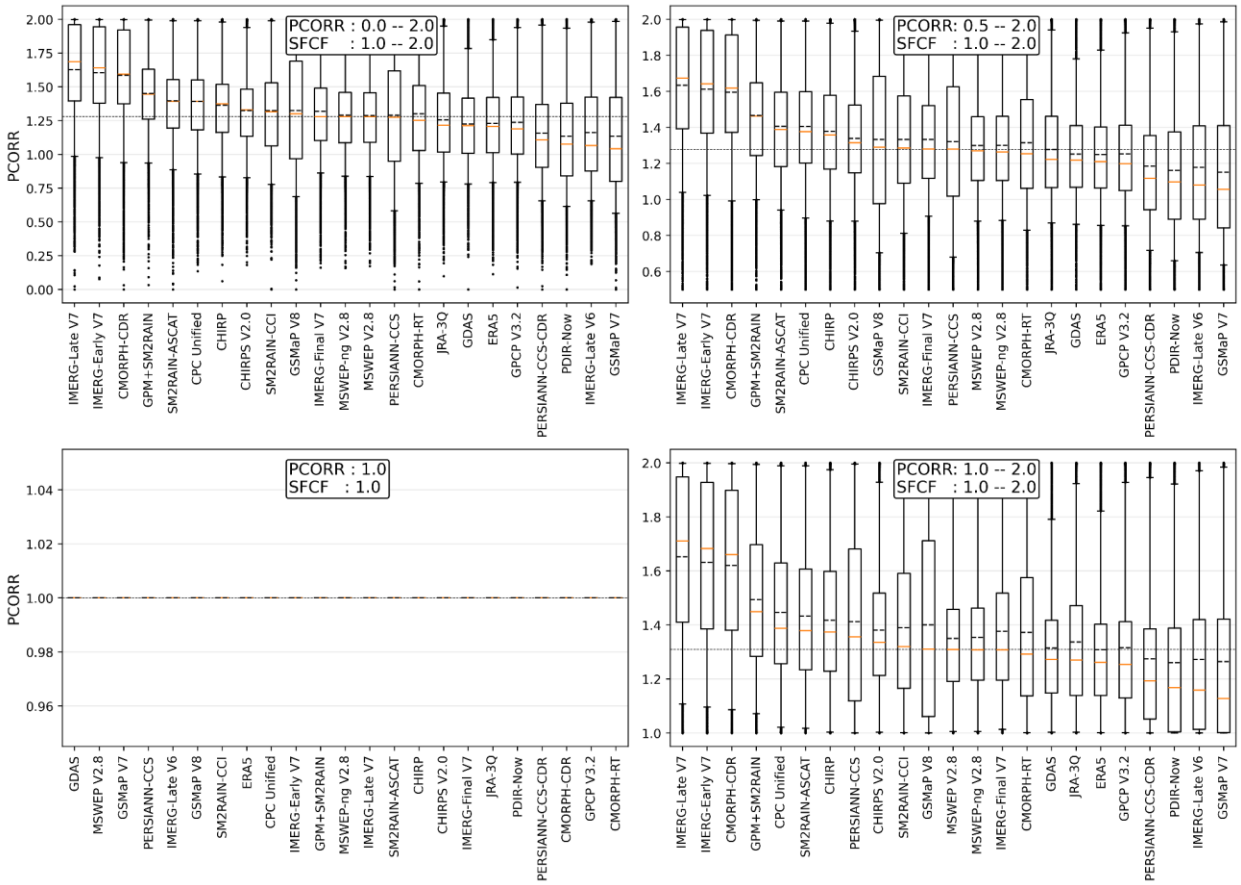

Fig S27. Comparison of distribution of calibrated PCORR values from four calibration scenarios.

We have added the following sentences in our manuscript to describe this.

Lines 137 - 144: *"To assess the influence of systematic P bias correction using the PCORR and SFCF adjustment factors on model performance, we explored four calibration scenarios with varying bounds for the PCORR and SFCF parameters. In the first scenario, PCORR was allowed to vary between 0.0 and 2.0, providing full flexibility to adjust for both under- and overestimation of P, while SFCF was allowed to vary between 1.0 to 2.0. The second scenario limited PCORR to the range 0.5–2.0, while keeping the range of SFCF between 1.0 and 2.0. The third scenario fixed both PCORR and SFCF parameters at 1.0, effectively disabling P bias correction. The fourth scenario constrained both PCORR and SFCF to the range 1.0–2.0, allowing only upward correction. These scenarios enabled us to evaluate the sensitivity of model performance to P bias correction and assess the robustness of P dataset rankings under varying calibration constraints."*

Lines 239 - 245: *"The overall ranking of P datasets remained largely consistent across the four PCORR calibration scenarios (Supplement Fig. S26). However, in the scenario where PCORR and SFCF were fixed at 1.0, GPCP V3.2 and ERA5 showed improved relative rankings—not due to higher performance, but because other datasets experienced greater performance drops under this constraint. Most datasets showed little sensitivity to the PCORR bound below 1.0, but a few—namely PDIR-Now, GSMaP V7, PERSIANN-CCS-CDR, and IMERG-Late V6—exhibited notable use of PCORR values below 1.0 (Supplement Fig. S27). This suggests that these datasets tend to overestimate P in certain catchments, and that downward rescaling improves their hydrological performance".*

The HBV model is applied in a lumped configuration, considering catchment-averaged forcing data. While this is understandable for a global-scale study, it would be very interesting to explore whether a semi-distributed configuration that accounts more explicitly for P gradients related to elevation, could provide additional insights, particularly in mountainous or topographically complex regions.

It would also be helpful to include a short statement in the Limitations Section acknowledging the assumption of constant land cover over the analysis period. Land cover changes can influence hydrological responses and might introduce some uncertainties in the model performance, especially over multi-decadal periods.

Response: Thank you for highlighting this important point. You are correct that in this study we did not subdivide each catchment into sub-catchments or sub-basins; instead, we used catchment-averaged forcing data to drive the HBV model. While using a semi-distributed model configuration with elevation bands would better account for spatial variability in precipitation and temperature—particularly in mountainous or snow-influenced catchments—implementing such an approach would significantly increase computational demand, especially at the global scale of this study. We therefore considered this beyond the scope of the current work, which focuses on providing a first large-scale comparative assessment of precipitation datasets. Nonetheless, we fully agree that exploring the use of elevation bands would be a valuable direction for future research to further improve hydrological simulations in complex terrain.

We also did not consider land-use and land cover changes with time which can occur when the simulation period is long. However, we do not think this will significantly alter the findings from this study i.e. regarding the suitability of P datasets for different regions. We have added both limitations to the 'Potential Limitations and Future Work' section of the manuscript.

Lines 343 – 347: "*Additionally, it does not account for spatio-temporal variations in land cover or use and relies on catchment-averaged meteorological forcings, omitting sub-catchment variability in climate and terrain. More complex (semi-)distributed models with hydrologic response units or elevation bands may yield improved simulations (Gu et al., 2023). However, we do not expect this to materially affect the relative performance ranking of the P datasets or the main conclusions.*".

Since PCORR and SFCF are calibrated, I also recommend including a few sentences explaining which types of biases are captured by the beta component of the KGE. This would clarify the references throughout the manuscript to over- and underestimation of datasets, which in part, could be related to biases of different magnitudes for specific P intensities, and the skill of the products to accurately detect P events.

Response: Thank you for this helpful comment. The β component of KGE reflects the ratio of the mean simulated to mean observed streamflow and thus captures systematic biases in the volume of simulated streamflow relative to observations. While PCORR and SFCF parameters of the HBV model are calibrated to correct for overall biases in precipitation and snowmelt inputs, residual biases may remain due to limitations in the precipitation product's ability to correctly represent the spatial and temporal distribution of P intensities and magnitudes (Sun et al., 2017). These residual biases can still influence streamflow and are partially reflected in the β term of KGE. We have clarified this in the revised manuscript.

Lines 161 – 164: "*While the PCORR and SFCF parameters, which account for systematic biases, were calibrated, the β component of KGE reflects residual biases that may persist due to limitations in the P dataset's ability to accurately represent the spatial and temporal distribution of precipitation intensities and magnitudes (Sun et al., 2017)"*.

In Section 2.2, the authors mention that streamflow records of selected catchments must span more than three years. Could the authors clarify if these years must be consecutive? Similarly, the rationale for requiring more than 10 non-consecutive P events is not fully explained. How was this threshold determined?

Response: One of the requirements for a catchment to be considered or not was to have at least 3 years total observation streamflow record. This record needs not to be consecutive.

Line 95: "*The total streamflow record had to be >3~years, not necessarily consecutive.*"

Additional minor suggestions:

Response: Thank you for your suggestions. We have implemented them in our updated manuscript.

Table 1: In the “Temporal resolution” column please use either “30 min.” or “30 min” consistently.

Response: Thank you for highlighting this issue. We have uniformly used ‘min.’ in Table 1 for temporal resolution.

L156: It would be helpful to report the median KGE for MSWEP V2.8 here for easier comparison with other products.

Response: The overall median KGE value for MSWEP V2.8 is 0.75. We have mentioned this value in that sentence.

Lines 176 – 177: “The multi-source MSWEP V2.8 dataset (Beck et al., 2019b) demonstrated the best overall performance (median KGE of 0.78).”

L189: Likewise, indicate the median KGE for CHIRPS V2.0.

Response: The overall median KGE value for CHIRPS V2.0 is 0.66. We have modified the sentence to include this value.

Lines 208 – 209: “. In contrast, CHIRPS V2.0 (median KGE of 0.66) applies five-day gauge corrections, while the other datasets apply monthly corrections, which provide fewer benefits at the daily time scale.”

Table 3: IMERG-Final V7 also performs best over tropical regions and should be marked in bold, as is done for MSWEP V2.8.

Response: You are correct that the median KGE values for both IMERG-Final V7 and MSWEP V2.8 in tropical regions were 0.66. However, we had highlighted only MSWEP V2.8 due to its coverage of a larger number of catchments compared to IMERG-Final V7. In the revised Table 3, we have now included the number of catchments for each dataset within each climate zone. The first row shows the total number of catchments, but since the availability of precipitation data varies across datasets, the actual number of catchments used for each dataset also differs.

Throughout the manuscript, “evaporation” and “evapotranspiration” are used interchangeably. Please consider clarifying and using these terms consistently.

Response: Thank you for highlighting this issue. We used the term “evaporation” instead of “evapotranspiration” based on the recommendation by Miralles et al. (2020). Additionally, Seibert and Vis (2012) also use the term “evaporation” when describing the structure of the HBV model. To maintain consistency, we have replaced the two instances of “evapotranspiration” with “evaporation” in the manuscript.

References:

Miralles, D. G., Brutsaert, W., Dolman, A. J., & Gash, J. H. (2020). On the use of the term “evapotranspiration”. *Water Resources Research*, 56(11), e2020WR028055. https://doi.org/10.1029/2020WR028055

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrology and Earth System Sciences, 16, 3315–3325, 2012.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, *56*(1), 79-107.