Review of EGUsphere-2024-4194

Title

Comprehensive Global Assessment of 23 Gridded Precipitation Datasets Across 16925 Catchments Using Hydrological Modeling

By: Ather Abbas, Yuan Yang, Ming Pan, Yves Tramblay, Chaopeng Shen, Haoyu Ji, Solomon H. Gebrechorkos, Florian Pappenberger, JonCheol Pyo, Dapeng Feng, George Huffman, Phu Hguyen, Christian Massari, Luca Brocca, Tan Jackson, & Hylke E. Beck

This manuscript provides an extensive evaluation of the hydrological performance of 23 gridded precipitation (P) datasets by calibrating a hydrological model over 16,295 catchments across the globe. The 23 P datasets are chosen based on their availability at (sub)-daily scale and their (quasi)-global coverage. Among them, 1 dataset is gauge-based only, 3 are reanalysis-based only, 12 are satellite-based only, 3 combine gauge and satellite data, 2 combine reanalysis and satellite data, and 2 combine gauge, reanalysis and satellite data. A conceptual hydrological model (HBV) is used to simulate daily streamflow and is calibrated with each P dataset using the evolutionary algorithm. The Kling-Gupta Efficiency (KGE) is used to assess the hydrological performance of each P dataset across 16,295 catchments.

Response: Thank you very much for your time to review our manuscript and your thoughtful comments. 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.

General Comments:

This study tackles a crucial issue in hydrological modeling, which is becoming increasingly important for many potential users who often lack guidance and confidence when navigating the wide range of products available for addressing specific hydrological problems across different geographical regions. The manuscript is well-written and well-structured; however, several issues need to be addressed to strengthen the robustness of the findings and conclusions.

Response: Please find the point-by-point response below.

The inclusion criteria stated in Section 2.2 (L80-104) are questionable because of their subjective nature and arbitrary setting. To ensure sufficient data for calibration, the number of events (defined as runoff > 5 mm d^{-1}) must exceed 10 non-consecutively (L96-97). However, some fixed values were set by the authors without providing a clear explanation of their rationale behind their selection. Would it be more appropriate to use a percentile of runoff instead of a fixed value for adaptation across catchments with different hydrological regimes? Similarly, to filter out catchments with erroneous streamflow and catchment boundary data, the authors set the mean annual runoff to be ≥ 5 and < 5000 mm yr^{-1} (L98-99). However, the range of mean annual runoff values can vary significantly across different climatic zones, with arid regions ranging from 0 to 100 mm yr^{-1}, while tropical regions can vary from 800 to over 2000 mm yr^{-1}. It would be appreciated if the authors could provide further explanation and justification for their inclusion criteria.

Response: Thank you for raising this important point. We agree that our inclusion criteria are subjective and may raise questions. However, for each criteria, we clearly explained the purpose it serves in the text. Additionally, choosing different criteria is unlikely to result in a different performance ranking of P datasets or different conclusions. Regarding the criteria of >10 events of >5 mm/d, this is to ensure we have sufficient dynamic events for calibration, and we think this criteria is not controversial, and is suitable for all climate zones, including arid ones. We are not convinced about using percentiles, as this might unrealistically identify very small streamflow occurrences as events, in case a runoff record is largely devoid of any runoff. Regarding the mean annual runoff threshold, in our experience <5 mm/yr or >5000 mm/yr of runoff is extremely rare, and therefore this criteria is useful to identify catchments with erroneous data, for example due to a mistaken unit conversion. We do not think this criteria is controversial or requires a longer justification than we already have.

The hydrological performance of P datasets with higher spatial resolution might be compromised when using catchment-mean P to drive the hydrological model, as the more detailed spatial information from these P datasets is lost. It is somewhat surprising to see that CPC Unified ranks second, given its coarse spatial resolution (0.5°). In addition, it is quite interesting that JRA-3Q performs better than its higher spatial resolution counterparts (ERA5 and GDAS). It is suspected that using catchment-mean P might mask the advantages of higher spatial resolution, leading to the conclusion that “higher spatial resolution does not guarantee better performance, especially when data is aggregated at the catchment scale" (L153-154). This may hold true for catchments dominated by a single climate or with relatively uniform topography, where spatial variability in precipitation has less influence. However, in mountainous, snow-dominated, or mixed-climatic catchments, the hydrological response cannot be adequately captured without detailed spatial P information. As a result, the true value of higher spatial resolution datasets may be underestimated, potentially biasing the selection of P datasets for hydrological modelling.

Response: Thank you for your comment. The usage of catchment averaged forcing data is one of the limitations of current work which can diminish the advantage of high-resolution P datasets over lower resolution datasets especially for large catchments. One way of overcoming this limitation is dividing the catchment into sub-basins, for example based upon elevation bands. 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, and we are quite confident it won't change the performance ranking of the P datasets and our main conclusions. We therefore considered this beyond the scope of the current work. Nevertheless, we have added this point in "Potential Limitations and Future Work" section of our 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.*"

The use of PCORR parameter to mitigate systematic biases in P datasets during calibration may present challenges because it adjusts only for P underestimation by setting the range between 1 to 2. This focuses on correcting underestimation without addressing P overestimation could disproportionately affect datasets prone to overestimation, potentially skewing performance evaluations. For instance, datasets like PDIR-Now and JRA-3Q, which experience overestimation, have low median KGE scores in some streamflow data sources (L277-280). It would be appreciated if the authors could provide a more comprehensive explanation and justification for their focus on mitigating only P underestimation.

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 reviewer 3. 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.

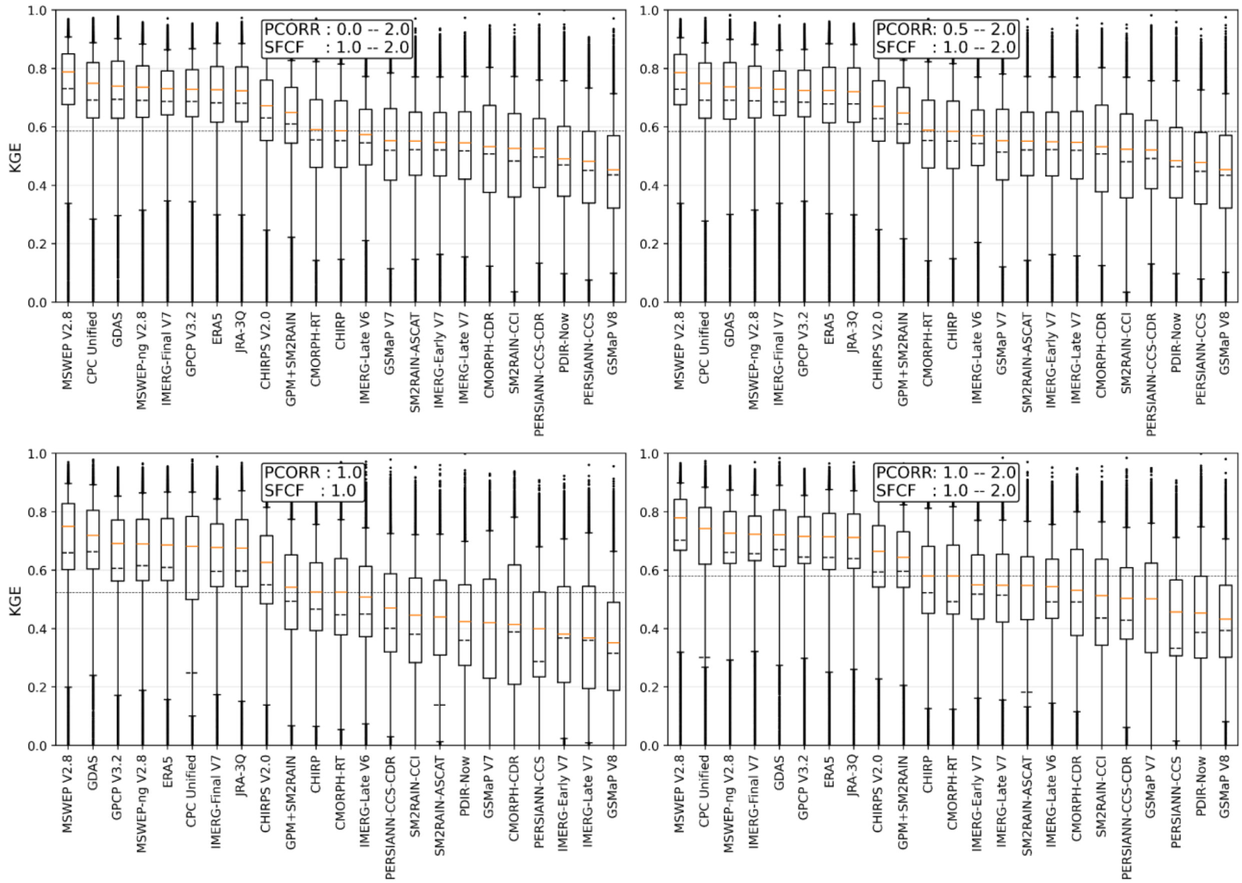

Fig. S26. Kling-Gupta Efficiency values for P datasets in different scenarios of PCORR and SFCF calibration.

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.

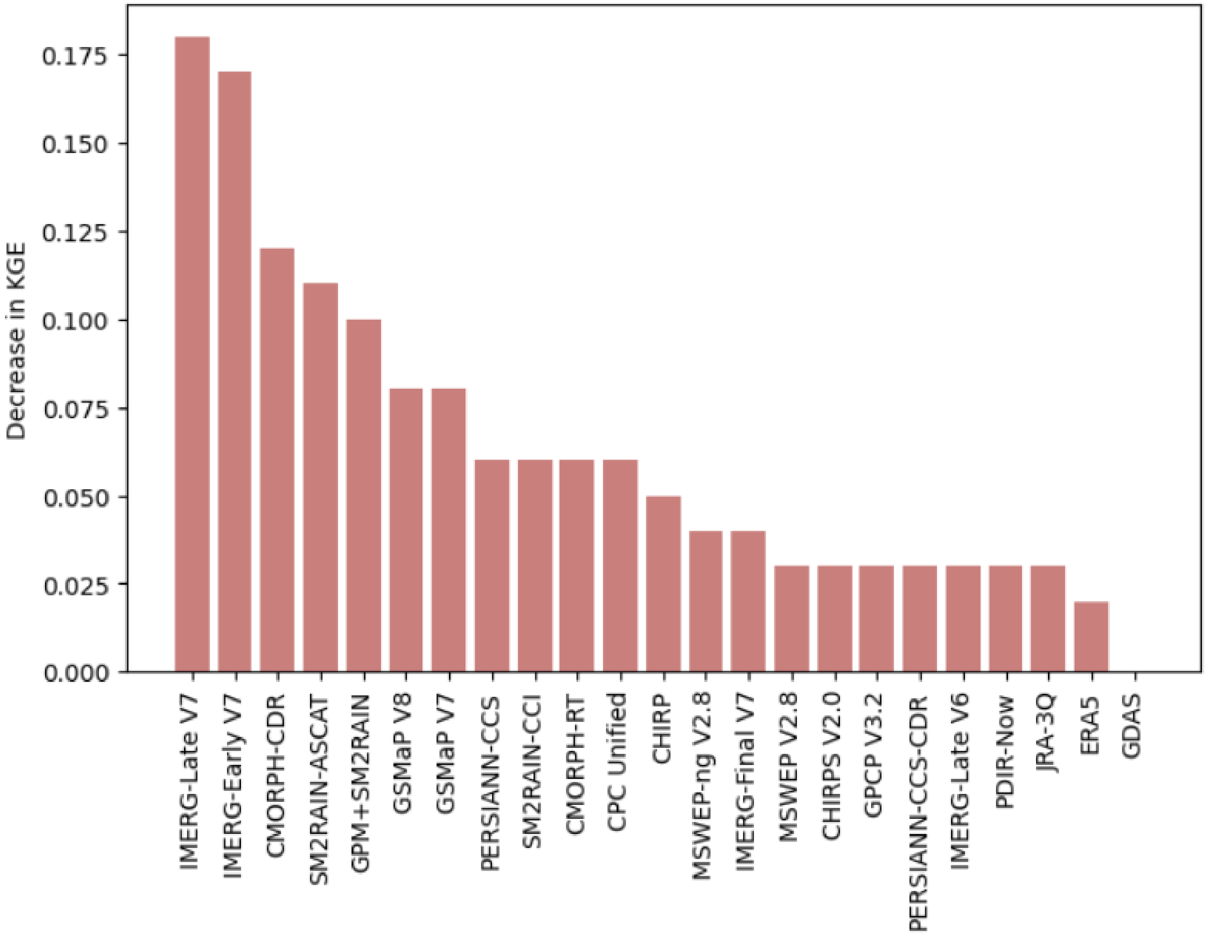

Fig. Decrease in KGE values when PCORR and SFCF were kept constant during calibration (1.0) as compared to when they were allowed to vary between 1.0 to 1.0.

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.

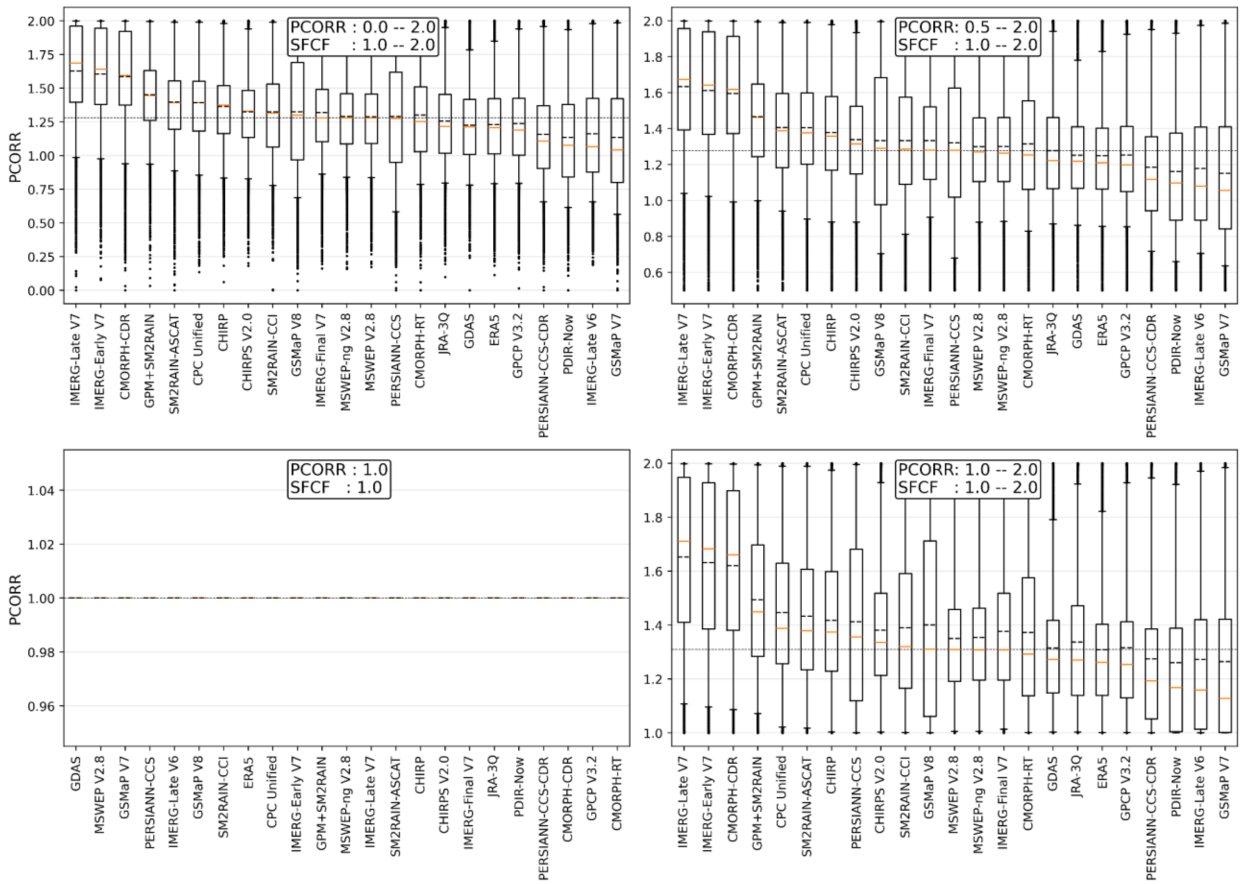

Fig S27. Comparison of distribution of calibrated PCORR values from four calibration scenarios.

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".*

Specific Comments:

L57-60: It would be appreciated if the authors could provide some basic information about the new datasets in the Gridded P Datasets section 2.1.

Response: Table 1 summarizes basic information of each of the 23 P datasets including data source (satellite, (re)analysis, gauge), spatial and temporal resolution, spatial and temporal coverage, time latency, full name as well as the reference.

L121-124: It is very unclear that how the model was initialized when 10 years of prior P data were not available. Did the authors just concatenate the same available P data n times to achieve the desired length? Or did the authors use any rainfall generators to produce stochastic P data? Please clarify and justify the use of "multiple times using the available P data until a total of more than 10 years was accumulated"?

Response: Yes, we concatenated the forcing data to achieve the desired length. This is because several years of initialization is necessary to bring the stores to optimum level. We would also like to highlight that even if we initialized the model by running it for 10 years, we still did not use the simulated results from the first 365 days to calculate performance during calibration. We have further clarified this in our manuscript.

L124-126: It would be appreciated if the authors could provide more information and description about the evolutionary algorithm.

Response: We used an Evolutionary Algorithm (EA) to calibrate the HBV model parameters. EA is a population-based optimization method where each individual represents a potential parameter set, evaluated using the Kling-Gupta Efficiency (KGE) and constrained by physically meaningful parameter ranges. The algorithm evolves the population across generations through biologically inspired operations such as crossover and mutation. In our implementation, 90% of the offspring were generated using blend crossover and 10% through Gaussian-based mutation. Offspring were then evaluated by running the HBV model with the new parameters. This process was repeated until a stopping criterion was met—either reaching the maximum number of generations or detecting convergence, defined as no improvement in the best KGE score greater than

0.0001 over 10 consecutive generations. We have added more description about it in the main manuscript.

Lines 136 - 139: "*We used a (μ + λ) evolutionary algorithm, which is a population-based optimization method that iteratively evolves solutions through selection, crossover, and mutation to maximize the Kling-Gupta Efficiency (KGE) objective. The algorithm was implemented using version 1.4 of Distributed Evolutionary Algorithms in Python (DEAP) library (Ashlock, 2010; Fortin et al., 2012), with a population size (μ) of 20, a recombination pool size (λ) of 48, crossover of 90% and Gaussian based mutation of 10%. To ensure convergence, the optimization process was terminated if the best KGE value did not improve by more than 0.01 for 5 consecutive generations after a minimum of 25 generations.*"

L128-132: For a particular catchment, the full period of overlapping streamflow and P data could be different because of the differences in the temporal availability of the P datasets. In this regard, will such differences also cause instability in the performance score?

Response: Thank you for this insightful comment. The temporal range of data for each catchment is different from each other. This is due to the difference in availability of streamflow records for different catchments and availability of P datasets. Unfortunately using a consistent time period for each P dataset would lead to exclusion of certain datasets from the analysis. However, since the number of catchments are very large, we believe the median performance metrics presented in our study for each region will not vary if a constant temporal range for each P dataset were used.

L170-176: Will the poor performance of PDIR-Now due to the inability of PCORR in adjusting overestimation of the P dataset?

Response: Thank you for your observation. This is partially correct in the case of PDIR-Now. We analyzed the effect of four calibration configurations for the PCORR and SFCF parameters in the HBV model, applied across all 23 P products and all stations. The table below summarizes the performance (in terms of KGE) and the median calibrated PCORR and SFCF values for PDIR-Now under each scenario:

Scenarios	KGE	PCORR	SFCF
PCORR: 0–2, SFCF: 1–2	0.49	1.08	1.39
PCORR: 0.5–2, SFCF: 1–2	0.48	1.10	1.4
PCORR: 1.0 (fixed), SFCF: 1.0 (fixed)	0.42	1.0	1.0
PCORR: 1.0–2.0, SFCF: 1–2	0.45	1.17	1.36

The performance improved in the first two scenarios (median KGE = 0.49 and 0.48, respectively) when PCORR was allowed to vary below 1.0. This added flexibility enabled the model to slightly reduce precipitation where needed, suggesting that PDIR-Now may overestimate precipitation in some catchments. Therefore, your observation is valid: the lower performance of PDIR-Now can be partially attributed to the restriction of PCORR to values ≥ 1.0, which limits its ability to adjust for overestimated precipitation. Allowing PCORR to vary below 1.0 partially improves model performance by enabling both upward and downward corrections. However, this pattern does not hold for all precipitation datasets. The figure below shows the improvement in KGE when the PCORR calibration range was expanded from 1.0–2.0 (scenario 4) to 0.0–2.0 (scenario 1). The most notable improvement is observed for GSMaP V7, with an increase in KGE of 0.06, followed by PDIR-Now, which shows an improvement of 0.04. For most of the remaining precipitation datasets, the change in KGE was marginal, with improvements generally around 0.01, indicating limited benefit from allowing PCORR values below 1.0.

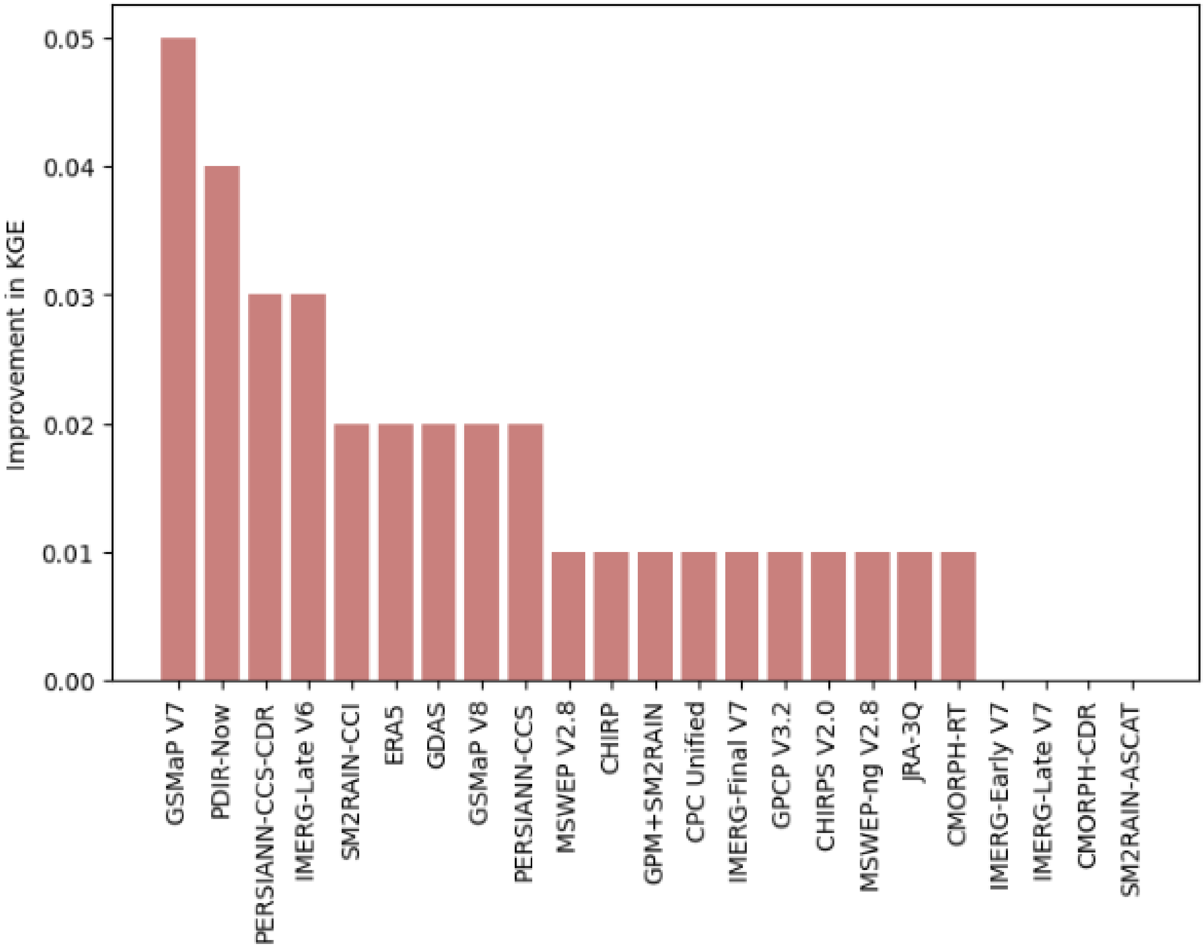

Fig S28. Improvement in KGE values for different P datasets by increasing the range of PCORR from 1.0-2.0 to 0.0 to 2.0 during calibration.

We have added the following lines in section 3.1 of manuscript

Lines 247 – 253: "*The lower performance of PDIR-Now can partially be attributed to the inability of PCORR in adjusting overestimation of the P. This is evident from lower calibrated values of PCORR when its value was allowed to vary below 1.0. The median calibrated PCORR value decreased from 1.2 to 1.1 which also resulted in a slight improvement in median KGE value from 0.43 to 0.47. Further analysis revealed that the largest decrease in median calibrated PCORR (1.0 to 0.7) and consequently improvement in KGE (0.15 to 0.37) was observed in CAMELS-GB (Supplement Fig. S29). However, a comparison of all P datasets indicates that improvement in KGE for most other P datasets was insignificant, affirming that the range of PCORR (1.0–2.0) is reasonable for most P datasets (Supplement Fig. S28)*."

L310-312: Could the authors elaborate further how the alignment of streamflow stations with meteorological network might favour gauge-based and reanalysis-based P datasets over satellite-only P datasets?

Response: This is primarily because gauge-based P datasets incorporate gauge data and reanalysis-based P datasets assimilate temperature, humidity, and other observations in these regions. Since streamflow gauges are typically located in regions with denser meteorological networks, these datasets tend to perform better in such areas. As a result, when performance is evaluated using streamflow data from these regions, it may lead to a slight overestimation of the relative performance of gauge-based and reanalysis-based datasets compared to satellite-only datasets, which do not benefit from local gauge input.

Remarks:

L54: typo "result n biased conclusions"

Response: Thank you for highlighting the typo. We have corrected this in the updated manuscript.

Lines 55 – 57: "*Additionally, several studies failed to re-calibrate the hydrological model for each P dataset, including the recent global assessment by Gebrechorkos et al. (2023), which could result in biased conclusions*."

L72: should it be "IMERG-Early V7" instead of "IMERG-Early V6"?

Response: Thank you. Yes, we have used IMERG-Early V7 and not IMERG-Early V6 in this study. We have corrected the sentence.

L275: please explain “TOVS-to-ATOVS transition. Thank you.

Response: The TOVS-to-ATOVS transition in ERA5 refers to a change in satellite observation systems. TOVS (TIROS Operational Vertical Sounder) was an older generation of sounders used from the 1970s to the 1990s. In 1998–1999, it was replaced by the more advanced ATOVS (Advanced TIROS Operational Vertical Sounder) system.