Comment: However, I think the annual indices with number of available daily/sub-daily data should be reported as 'NaN' or removed. If we have both Year A and Year B has the same number of daily data, e.g., ~180 days, to derive the annual indices. But Year A is available from Jan to Jun, while Year B is available from Jul to Dec. Such inconsistency will result in bias for calibrating/validating model simulation because they represent streamflow characteristic from different seasons. Therefore, although the users can decide if a year is "good" or not, they will not know if all the good years are consistent in the time period.

Thank you for this comment. We acknowledge that the temporal distribution of available data throughout the year can influence analysis and interpretation, thus **we added a new field "month with nan>10 days" in the yearly indices table, which includes the list of the months with over 10 days of NaN measurement (see examples from the last columns of Tables R1 and R2).** We did not set the annual indices as "NaN" directly for years with a fixed number of available daily observations because the purposes of data users can be very different, and we would like to make the selection criteria as flexible as possible to allow for more potential utilizations. For instance, Table R1 shows a gauge where AMF usually occurs in summer. Therefore, despite that data for 2001-2005 are not intact, the AMF indices are generally reliable, while mean, median and low flow data are not usable. For another case in Table R2, although observation days exceed 250 in 1988 and 2021, streamflow data for consecutive three months are missing, thus annual means of these two years are still not equally representative as other years. We hope the additional field of the data table satisfies the quality control while leaving flexibility. We've revised our dataset in the uploaded file.

Year	mean	maximum (AMF)	AMF occurrence date	number of days with Q=0 (days)	valid observation days (days)	month with nan>10 days
2001	49.48039	691	['2001/07/29']	0	181	[1, 2, 3, 4, 11, 12]
2002	49.65326	795	['2002/06/28']	0	184	[1, 2, 3, 4, 11, 12]
2003	153.8382	1115	['2003/08/16']	0	185	[1, 2, 3, 4, 11, 12]
2004	98.21307	1579.714	['2004/07/17']	0	185	[1, 2, 3, 4, 11, 12]
2005	163.0907	940.25	['2005/09/28']	0	245	[1, 2, 3, 4]
2006	38.06584	715.25	['2006/07/05']	0	364	[]
2007	50.60067	820	['2007/07/14']	0	365	[]
2008	32.33534	468.4	['2008/07/24']	0	366	[]
2009	83.40167	909.3	['2009/06/19']	0	365	[]
2010	192.2253	4474.231	['2010/07/25']	0	365	[]
2011	40.71113	616.4815	['2011/09/15']	0	365	[]
2012	50.16656	335.5769	['2012/07/06']	0	366	[]
2013	23.58099	214.875	['2013/07/02']	0	365	[]
2014	12.67763	44.52917	['2014/10/02']	0	365	[]
2015	23.53541	123.16	['2015/05/04']	0	365	[]

Table R1 Selected from 62011800_China.csv.

year	median	mean	maximum (AMF)	AMF occurrence date	number of days with Q=0 (days)	valid observation days (days)	month with nan>10 days
1985	0	1.727559	126.307	1985/11/3	307	365	[]

1986	0	0.034907	6.177	1986/7/5	354	365	[]
1987	0	11.62648	775.377	1987/2/15	298	365	[]
1988	0	4.036762	230.29	1988/4/3	199	273	[10, 11, 12]
2005	0	0.038632	3.364	2005/6/23	200	212	[1, 2, 3, 4, 5]
2006	0	3.711151	488.019	2006/4/6	302	365	[]
2007	0	3.464066	338.672	2007/1/22	328	365	[]
2008	0	0.63915	81.378	2008/11/27	345	366	[]
2009	0	36.35345	1330.387	2009/1/9	277	365	[]
2010	0	6.590668	878.593	2010/1/8	285	365	[]
2011	0	4.40994	189.484	2011/1/18	250	365	[]
2012	0	0.789429	62.876	2012/3/16	328	366	[]
2013	0	0.138811	21.486	2013/11/29	345	365	[]
2014	0	2.281156	233.295	2014/3/1	311	365	[]
2015	0	2.899764	252.943	2015/12/28	335	365	[]
2016	0.0405	14.60942	589.812	2016/3/12	149	366	[]
2017	0	1.908384	145.384	2017/1/16	316	365	[]
2018	0	4.594542	673.417	2018/3/5	333	365	[]
2019	0	15.76645	1654.957	2019/3/28	268	365	[]
2020	0	2.696989	157.835	2020/1/27	315	366	[]
2021	0	1.02686	47.653	2021/2/13	211	250	[10, 11, 12]

Table R2 Selected from 001202A_BOM.csv.

Comment: I appreciate the authors' efforts to validate the watershed delineation based on my comment. I believe the contributing area should be reported for each gauge, at least from USGS. For example, the author can find the drainage area at this USGS gauge: hcps://waterdata.usgs.gov/monitoringloca Won/07374000/#parameterCode=00065&period=P7D&showMedian=true. I understand that GRDC gauge coordinates may be highly uncertain, so USGS gauges could be a good benchmark, which is at higher quality.

Thank you for your comment. We have actually validated the GSHA gauges against the HYDAT, GRDC, BOM, and USGS gauges and added the verification flag field in the dataset, but only the validation scatter plot of the previous three agencies were shown in the appendix. Here in Figure R1 we show the validation result of the USGS gauges. Correlation coefficient is 0.905 before removing the mismatched watersheds, and 0.999 after removing the mismatched watersheds. Based on the good match, we added the validation results of USGS areas in Figure B1 (Figure R2 in this reply file) in Appendix B in the latest revised version of the manuscript.

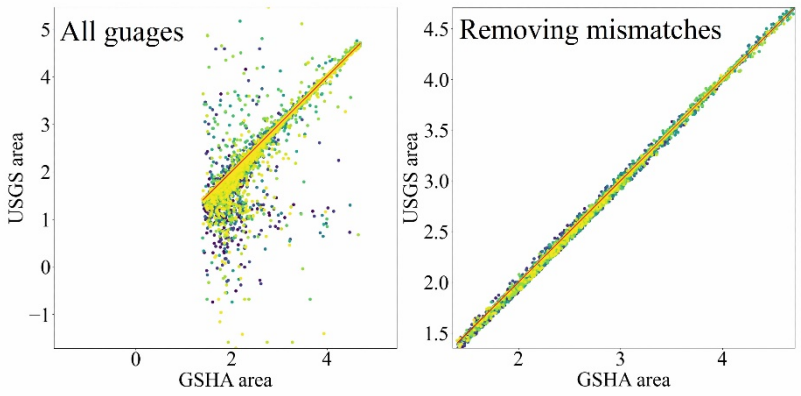

Figure R1 Validation of GSHA with officially reported areas of USGS gauges. Subfigure on the left is the result before removing the mismatched watersheds, and subfigure on the right is the result after removing the mismatched watersheds.

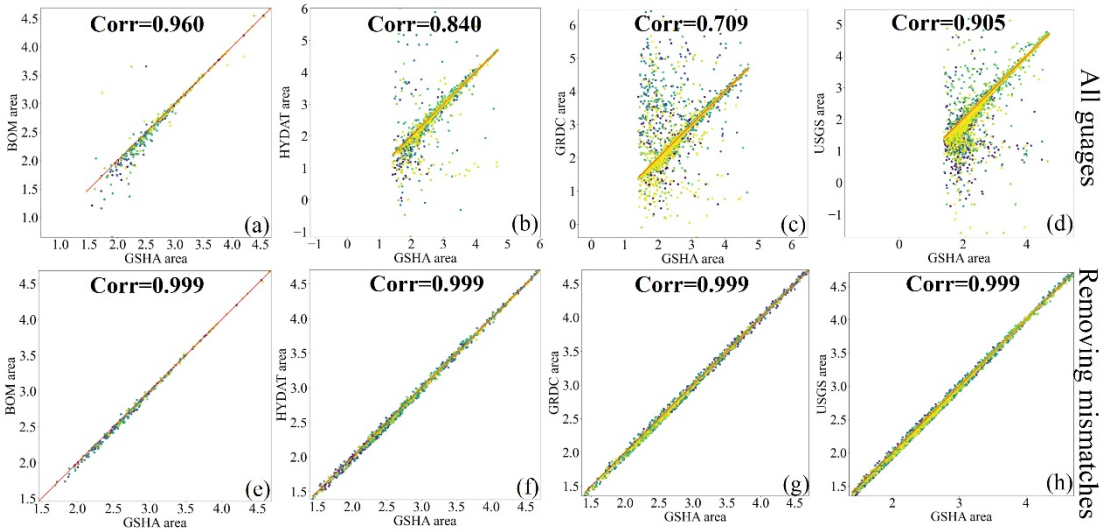

Figure R2 Validation of GSHA with officially reported areas of BOM (a, e), HYDAT (b, f), GRDC (c, g), and USGS (d, h). Subfigures (a) to (d) are the results before removing the mismatched watersheds, and subfigures (e) to (h) represent results after removing the mismatched watersheds. The Pearson correlation coefficient are represented by "Corr" in the figure. The areas are represented by the unit of (log10 km^2).

Comment: Do you mean for the case Xmin = 0 and Xmax > 0, the $\bar{X} = \frac{\text{Xmax}}{2}$?

Yes. In our cases except for temperature, Xmin = 0 and Xmax > 0, thus $\bar{X} = \frac{\text{Xmax}}{2}$ and uncertainty equals to 200%. For temperature, both Xmin and Xmax > 0, and uncertainty is smaller than 200%.

Do you assume the variable X varies linearly from Xmin to Xmax?

We assume X as linearly varied because we have two or three datasets (sample numbers in X) for the meteorological variables, and indicators such as interquartile range, STD, or CV cannot be calculated. We do not consider the distributions of measurement samples among these datasets, and used Xmax-Xmin as an alternative to the interquartile range/STD/CV in the calculation of standard measurement uncertainty [1]. Our estimate was in accordance with, and supported by

the concept of measurement uncertainty, where the actual true value is fixed, and the uncertainty range is acquired by the intervals of all measurement samples to represent the precision of a measuring system [2]. As the actual true values of the variables are unknown, we assume the mean of the datasets to be the alternative true values. The actual true values are assumed to lie within the range of the maximum and minimum values of the datasets (Xmax and Xmin) with a linear possibility.

Reference:

[1] White GH. Basics of estimating measurement uncertainty. Clin Biochem Rev. 2008 Aug;29 Suppl 1(Suppl 1): S53-60. PMID: 18852859; PMCID: PMC2556585.

[2] Libretexts. (2023, August 27). 1.3: Measurements, uncertainty and significant figures. Physics LibreTexts. https://phys.libretexts.org/Courses/Georgia_State_University/GSU-TM-Physics_I_(2211)/01%3A_Introduction_to_Physics_and_Measurements/1.03%3A_Measurements_Uncertainty_and_Significant_Figures#:~:text=Discrepancy%20is%20the%20difference%20between%20the%20measured%20value,then%20the%20discrepancy%20of%20the%20values%20is%20high.

Comment: Except the location and id, the contributing area can be used as the third criterion for paring the gauge in both GSIM and GSHA. Specifically, if the contributing area are not the same, there is a high probability that not the same gauge is used in GSIM and GSHA for comparison.

Thanks for the suggestion. We have already considered the area difference in our matching process, and we have actually mentioned “the GSIM gauge with a minimum distance and watershed area difference ≤5% to a GSHA gauge was considered” in our original manuscript (Line 377-378).

In addition, I don’t understand how the time step impact the annual streamflow indices, e.g., p90 that is reported in Figure R2. Should the estimate of the annual streamflow indices have based on 365 or 366 daily streamflow (if the data is available for the whole year)?

In our calculation, it was not required that streamflow observations are available throughout the year (365 or 366 days), and that observations start and end at 31st Dec. The percentiles were based on the available observations, since we gave several fields for data filtering purposes, including number of days with Q=0, valid observation days, and month with nan>10 days. For instance, for a summer monsoon-controlled watershed in Asia, if NaN values are concentrated in DJF, the p90 and AMF indices are not likely to be influenced. Therefore, it is possible that the missing values and inconsistencies of time span cause some discrepancies in GSIM and GSHA indices in Figure R3 (Figure R2 in last version of reply), but we do not require that the estimate of annual streamflow be based on 365/366 days of data.

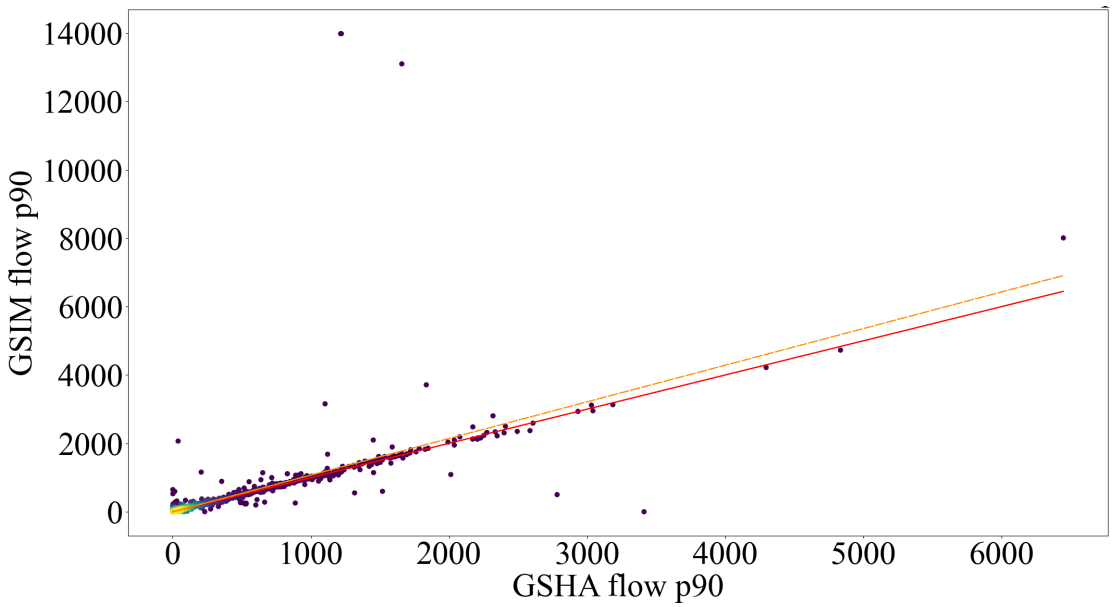

Figure R3 Validation of GSHA with GSIM streamflow 90 percentile. The red line is the 1:1 line, while the orange dotted line is the fitting line of the scatter points.