We thank the reviewers and the editor for handling our paper and offering constructive comments. Below please find our point-by-point response.

Reviewer #1

Comment 1: It would be very helpful to include stream order, COMID and NextDownID based on the MERIT hydrography. The thinking here is that many catchments are nested because the dataset is gauge-based. Including the suggested columns/fields could enable the users to incorporate upstream-downstream relationship in their analyses, for instance, where scaling behavior is key.

Thank you for your suggestion. We added another file "GSHA_MERITinfo.csv" in the "/Global_files/" directory with COMID, order, and NextDownID fields from the MERIT database. We also included the uparea field from MERIT as a comparison of our watershed boundaries.

Comment 2: Minor comment: the writing mixed present and past tenses, causing confusion on which actions are taken by the authors and which are taken in previous studies. For instance, the sentence from L288 to L292 is grammatically ill and hard to understand.

Thanks for your comment. We checked the verbs in the article and modified the tense of the wrong sentences. The description in 3.4.1 Meteorology datasets section involves information of many dataset names and might be confusing. Therefore, we modified the entire section to shorter sentences to clarify the descriptions. We also modified some other grammar mistakes in the text. We are sorry for the inconvenience.

Comment 3: The dataset involves merging several datasets – taking weighted average. I expect a brief introduction of the motivation/justification of applying certain method, for instance, Lu et al.,2021. This is important because it actually matters whether to merge several data sources or to simply provide their individual values. To me, merging is only meaningful when estimation error from individual merged component is informed. This is why the authors need to better justify/explain the merging of multiple datasets.

Thank you for your comment. We did not use the weighted average to merge the datasets, which is different from the purpose of Lu et al., 2021. The area-weighted approach was used to extract grid data to watersheds. Because the grids are fully or partially contained by the watershed boundaries, it is not accurate to simply calculate the arithmetic mean of all the grids intersecting with the watershed boundary. Therefore, we calculate a weight for each grid representing the proportion of grid contained by the watershed boundary. This follows the method adopted by Addor et al., 2017, which is the foundation of CAMELS datasets. For dataset merging, we agree that simply merging the datasets by their mean values without further estimating errors is irresponsible. Therefore, we treat each dataset as an independent estimation of the

variable and did not merge them for an averaged value. We added “based on the proportion of the grid area contained in the basin boundary” in line 348 to clarify the spatial aspect of this issue.

Reviewer #2

Major Comments #1

Line 163 – Line 170: The authors should clarify how they aggregated the daily streamflow to annual streamflow. There could be missing data for the streamflow in any year, and the missing days and the number of the missing days are not the same in different years. How the authors addressed the data gaps in the daily streamflow? For example, is there a criterion for the number of available days in a year that used to filter “good” years?

Thank you for your comment. We aggregated the daily streamflow by calculating their annual indices, such as annual mean, maximum, percentiles, as well as temporal characteristics such as maximum flood occurrence date, duration of high and low flow events. Therefore, we describe our data as streamflow characteristics instead of annual streamflow. **The dataset includes a “valid observation days” field**, which describes the number of days with available daily streamflow in the corresponding year, **as well as a “Q=0 days” field** representing the number of days with runoff measurement equal to 0. The data were not filtered or selected based on any criterion set by the authors, because **we would like to let the users decide how many available or non-zero measurements define a “good” year to them based on their research purposes and scales.** To make this clearer, we modified the sentence in lines 170-171 to “We also include numbers of zero observations and valid samples to allow flexible data screening by the users.”

In addition, I think monthly streamflow indices are more useful for modelers to calibrate and validate models. For example, previous studies have used the monthly time series from GSIM to calibrate large scale hydrological models and Earth system models.

Thanks for the comment. We now publicize the monthly indices of gauge observations (except for some transboundary watersheds) by calculating the monthly mean, maximum, percentiles, max flow occurrence date, number of days with Q=0, and valid observation days of the watersheds after 1979, and **attached the files can be found at https://zenodo.org/records/10127757**.

Major Comments #2

There is a lack of validation of watershed delineation. The watershed delineation could be one of the most important characteristics of GSHA, as many other variables were

extracted based on watershed boundary. I think the flow directions may be carefully validated in previous study, but it is important to validate the delineated watershed boundary. For example, most gauges reported watershed boundary or drainage area, which can be used as benchmark.

Thanks for your comment. Since we found out we do not have access to officially reported areas of all watersheds from agency websites, we validated our watershed areas for Australian Bureau of Meteorology 2022 (BOM), Canada National Water Data Archive 2022 (HYDAT), and The Global Runoff Data Centre 2022 (GRDC). The validation results are plotted in **Figure R1**.

There are indeed mismatches between GSHA areas and the officially reported areas by the agencies. As we used the MERIT Basins (Lin et al., 2019) for watershed dissolving, we do not question the sub-watersheds used in our delineation. After we compared our watershed area with officially reported area, it is found that some mismatches might occur when the gauge appears in the vicinity of the intersection point of a river reach and its main stream, which makes it difficult to decide which reach the gauge belongs to while matching the gauge to the MERIT river network. This explains why in Figure 1 most of the mismatches appear at relatively small areas.

To address this issue, we set the criteria of mismatched watershed as: (1) the area difference being over 20% of the officially reported area, and (2) the area ratio being over 10 times. **Under this criterion, 1.9% of BOM watersheds, 4.7% of HYDAT watersheds and 8.9% of GRDC watersheds are mismatched**, as plotted in **Figure 1** (a) to (c). After removing these watersheds, (d) to (f) show very good match of the watershed areas, with correlation coefficients reaching 0.999, suggesting that the remaining watersheds match with the officially recorded areas well.

Therefore, we decide to make the following modifications:

(1) **Add the area validation** to 3.7 Validation, 4.1 Technical Validation and Appendix B sections to inform the readers of the issue and the casual factors;
(2) **Add a flag field in the watershed list** of the dataset to describe the watersheds as “unverified”, “verified match”, and “verified mismatch”. As we do not have access to all official watershed areas, to simply remove the mismatched watersheds or to modify them might put the samples in the dataset under an unfair standard.
(3) We do not identify our mismatched gauges as “wrong” because whether our delineations are incorrect remains to be investigated, and after further check with other sources we will update our conclusions in the next version of GSHA. At the end of October, approximately 400 GRDC gauges updated their coordinates and some of them have experienced major deviations. After inquiring GRDC on this issue, we received their response as “*We occasionally receive updates on the metadata from the National Hydrological Services (NHS). This explains the smaller deviations, as it includes updates of longitude and latitude, as well as altitude and the size of the catchment area. The larger deviations occurred for the following reason. While*

recalculating the station-based catchment areas, there were some stations for which an exact derivation was not possible. The reason for this was that the coordinates were incorrect or inaccurate". This suggests that data from some agencies come with errors and uncertainties, and we will follow up on these updates to obtain more accurate information.

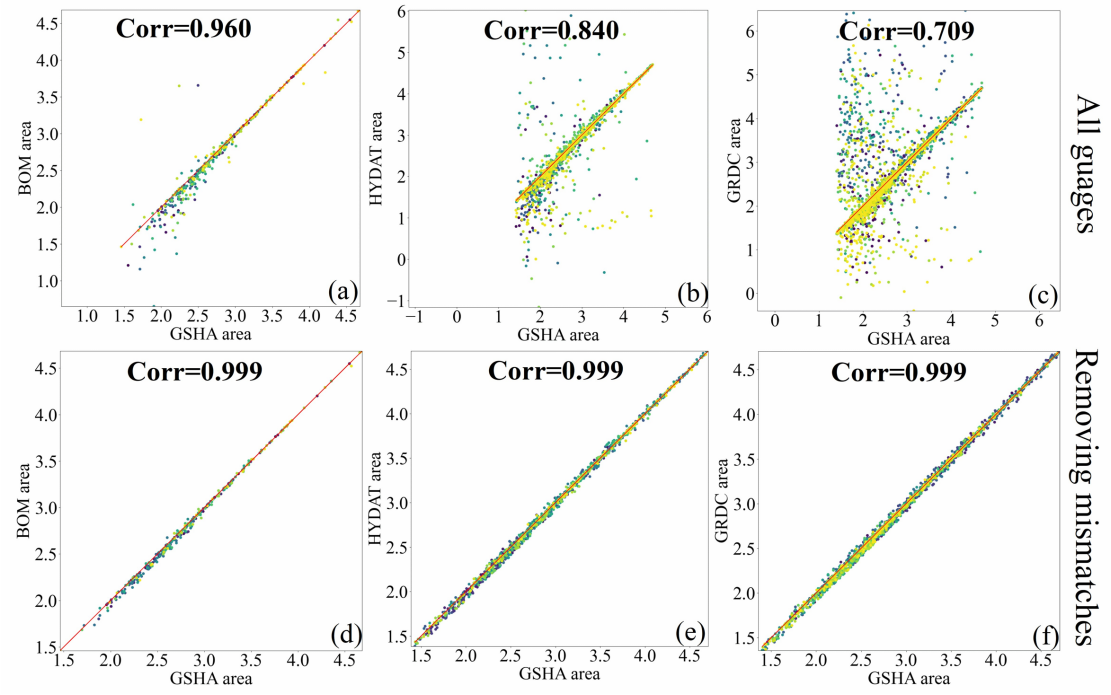

Figure R1 Validation of GSHA with officially reported areas of BOM (a, d), HYDAT (b, e), and GRDC (c, f). Subfigures (a) to (c) are the results before removing the mismatched watersheds, and subfigures (d) to (f) represent results after removing the mismatched watersheds. The Pearson correlation coefficient are represented by "Corr" in the figure. The areas are represented by the unit of (log10 km^2).

Major Comments #3

One of the novelties of GSHA comparing to existing large sample hydrology datasets is GSHA provide the uncertainty analysis for the selected variables. But I think current description of uncertainty estimate is not clear, and the method is not comprehensive. Specifically, Line 354-Line358: I don't think Eq (1) represents the uncertainty of the meteorological variables. As X_{max} and X_{min} represent the maximum and minimum values of the extracted variables fromeach individual dataset, X_{max} and X_{min} is more linked to the natural variability instead of uncertainty of that dataset. For the example of temperature, if we have a dataset give us X_{max}= 35°C, X_{min} = -5°C, and X = 10°C. Is the uncertainty of this dataset being 400%? And please further explain why the range of Eq (1) is between 0 and 200%.

However, based on the results in Figure 6, I think X_{max} and X_{min} are derived from all the datasets? I think the authors should further clarify the definition of uncertainty. In addition, X_{max} and X_{min} cannot capture the uncertainty in the temporal variability. It is possible for two datasets capture exact the same X_{max} and X_{min}, but have different distribution. Thus, X_{max} and X_{min} is not a good metric for analyzing the uncertainties from different datasets. I suggest the authors to include more metrics in the uncertainty analysis.

Thanks for the comment. The uncertainty we calculate represents the discrepancy between long-term means of the datasets, instead of the differences of each value in the time series. The X_{max} and X_{min} values are the maximum and minimum values in the **dataset ensembles (in our dataset two to three members included), rather than the max and min values in the temporal series**. We use this estimate to represent uncertainty of the mean value. Therefore, the distributions and variances inside each dataset are not considered. We understand that uncertainty should be represented by a range around the true value of the variable, but we do not know the true values of each variable at each particular date, and daily estimates from the datasets can be very biased. Therefore, we believe uncertainty range represented by discrepancy of the long term mean can be more meaningful compared to a time series of daily differences. 200% uncertainty occurs when one dataset $X_{min} = 0$ and $X_{max} > 0$. As we use K as temperature unit, there will be no negative value in the data.

To clarify this concept, we modified the sentence in line 368-371 as "We also provide uncertainty estimates of the meteorological variables by calculating the long-term mean of each dataset in each watershed, where the discrepancy between the maximum and minimum among the data sources (X_{max} and X_{min}) as a percentage of their mean ($\bar{X}$) was used in the uncertainty estimation".

Specific Comments

Line 160: upstream drainage ~~basin~~ area

Thank you for this comment. We modified the mistake and checked all descriptions of watershed area.

The section numbers from 3.2 to 3.7 were wrong.

We changed the wrong section numbers to 3.3 to 3.8.

Line 212: Need to define "shorter record length" explicitly.

We changed "shorter record length" to "fewer years of measurement" in line 215.

Line 216: CHP is defined in Table 3. But I think it is better to give the full name in the main text as well.

We added the full name of CHP at its first appearance in Line 219.

Line 318-Line 325: How you match the dams in GeoDAR to GSHA? Is it possible for a watershed to have several dams? How about the watershed that doesn't have a dam from GeoDAR?

We used the reservoir polygons in GeoDAR instead of the dam locations. To clarify the extraction process, we added the sentence "For reservoirs, we used the reservoir polygons in GeoDAR, which are spatially joined to GSHA watershed polygons. All the intersected reservoirs were considered contributory to the management of the corresponding watershed, and were used to calculate the total reservoir storage capacity and degree of regulation" in lines 351-354. For watersheds with multiple reservoirs, the sum of the capacities of the reservoirs were calculated. For watersheds with no reservoir, the capacity and DOR fields were set as empty. We manually checked a portion of the spatial join, and found the automatic spatial join approach to be reasonable.

Table 4: I think MSWEP is at spatial resolution of 0.1° × 0.1°. Please double check.

The version 2 of MSWEP is 0.1° × 0.1° resolution. However, our attempt of this research started before 2019. Therefore, MSWEP v1 with a 0.25° × 0.25° resolution were used. We will extract MSWEP v2 values in the updated version of GSHA in the near future.

Figure 5a, b, and c: Are the X and Y axis normalized? I suggest the authors to plot the original data (e.g., in m3/s) to demonstrate that no system errors were introduced during the processing of GSHA. In addition, the authors should explain why the comparison of some watersheds are very off from the 1:1 line. In my understanding, both GSIM and GSHA were derived from gauge observations for the streamflow indices. Therefore, same gauge measurements should be used at the same watershed in both GSIM and GSHA. Is the significant difference caused by (1) different gauges were used, or (2) different method was applied to address the data gaps in the gauge measurements (see my **Major Comments #1**), etc. Overall, I think it is useful for the authors to further explain the significant discrepancies in those gauges.

Thanks for this comment. The X and Y axes are log10 results of the original data, since the original data plot can be dominated by a few very large observations, as shown in **Figure R2**. Therefore, in order to clearly show the distribution of the majority of the samples, we used the log10 of original data. We are sorry for not clarifying this in the text. We added "The unit of X and Y axes in (a), (b). and (c) is long10 m3/s" in the caption of Figure 5.09.

We matched the gauges by their latitudes and longitudes, each point should represent the pair of the same gauge. However, the location matching might confuse a small proportion of very close gauges. Therefore, it is possible that the different gauges used cause deviations of validation results, and **we think locational error is the most**

significant factor causing the problem. However, currently we do not have a proper method to find out which gauge pairs are wrong based on ids and locations, thus we plotted all pairs in the validation figures. For data selection, GSIM suggested that "Given that data quality requirements can vary substantially, it will remain the work of individual users to establish selection criteria for each study, thereby finding a trade-off between data quantity (number of gauges) and data quality (record length, missing periods)" (Gudmundsson et al., 2018), which is consistent with our decision not to filter the observations as mentioned in the reply of Major Comments #1. However, according to the time step in the GSIM file, the first time step and last time step are usually 31st Dec., apart from some missing values, while we did not process our data that way. This might cause some discrepancies, but with monthly indices provided, we believe more accurate analysis can be carried out. We added these two reasons in the 4.1 Technical Validation section to inform the readers of these causes of differences.

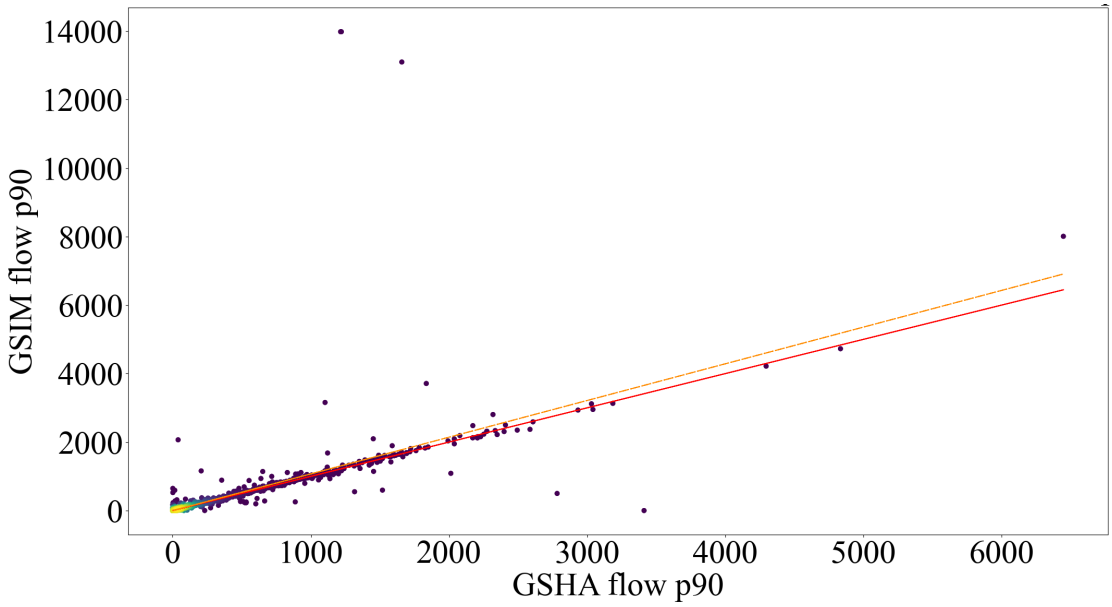

Figure R2 Validation of GSHA with GSIM streamflow 90 percentile. The red line is the 1:1 line, while the orange dotted line is the fitting line of the scatter points.

Figure 7: I don't think the decline of uncertainty as the watershed area increases is obvious for longwave radiation.

We agree with this comment and modified the description as "The most obvious decline comes from ET (green), which is highly dependent on the land surface conditions and are significantly affected by land surface spatial heterogeneity, thus benefiting the most from spatial averaging for large river basins. Longwave radiation uncertainty (red) experiences a moderate decline, mainly due to its connection with land surface complexity and cloud conditions" in lines 483-486.

Line 517 – Line 529: The analysis of runoff coefficient and its changing trend in the past few decades is very interesting and is very critical for us to understand response of

hydrological cycle to global warming. Such analysis with observed streamflow is more convincing than model simulations, which can be highly biased. I believe there exist some other studies focusing on this topic, such as runoff trend in this historical period. I wonder if the authors can give more discussion for this analysis and include more references.

Thanks for the comment. We added some discussions on runoff coefficient (RC) analysis considering land cover change in section 4.4, which are largely regional studies or focusing on individual cases. Additionally, our investigation suggested that such analysis incorporating water consumption and other human modifications, especially on large scale, are still insufficient. Therefore, we believe there is still a gap on the identification of large-scale patterns of RC trend and its attribution. We will follow up on this topic and try to identify signals of water availability change and the casual factors based on observations.