Please see my answers in red.

Dear Anonymous Referee,
We greatly appreciate your constructive and enlightening comments, which has helped us substantially improve our dataset. We have expanded our dataset by including stations of other available databases, i.e., USGS, HYDAT, ANA, CCCRR, and BOM. Currently, the dataset covers 41263 stations. The manuscript has also been revised by adding contents about data merging and formatting, and comparative analysis. We believe the revised dataset and manuscript will be satisfying. Below are our point-to-point replies.

1. Databases. You merged the GRDC, WRIS, ArcticGRO, and CHY databases. Why did you not use many of the other sources that are available, such as USGS, HYDAT, ARCTICNET etc.? Your selection of databases resulted in a total of 9171 stations with daily data vs. 35002 stations in GSIM. On top of that, only 5548 timeseries could be used for the calculation of indices. That is a huge difference and analyses will produce different results on a global scale.
Response: Thanks for your question. We have expanded our dataset and included as many databases as possible. Some databases are not included for various reasons. For example, A Regional Hydrographic Data Network for the Pan-Arctic Region (ARCTICNET) is not included as the data have not been updated for a long time and are outdated with the latest records at around 2001 (Lammers et al., 2016, 2001). European Flow Regimes from International Experimental and Network Data (EWA) are not incorporated since another database we have included has integrated the database. Eventually, our new dataset covers 41263 stations, which is larger than GSIM. See L115-170 for details.

Great!

2. Databases: merging and formatting. You do not describe how you merged and formatted the timeseries of the different databases. Did you check that there were no duplicate stations across databases? How did you handle different data formats, e.g., did some data come with flags and if yes did you include them? How did you check for and merge metadata, e.g., did you check gauge locations, was catchment information included?
Response: Thanks for your questions. We have added relevant sentences to describe the merging and formatting. The duplicate stations have been identified and removed (see L128-137). Flags have been attached to every record, station, and yearly index according to specific rules. Original flags from databases have been translated into the standardized flags. (see L184-208 for details). The metadata are merged according to the fields of our dataset's fields (see L138141 and Table 5). We did not check the gauge locations since there is no way to judge whether one station's location is right or wrong. As for the catchment information, we have included the catchment information that databases provide.

OK, one common error in gauge location data is too much rounding of the coordinates' decimal degree, which could be easily tracked. Other methods of checking the location would be to inspect if points fall close to a river using satellite imagery. However, I appreciate that this last step is very time-intensive and beyond the scope of this work. But you could add a sentence to indicate that this was not done.

3. Quality checks of the streamflow indices. Your quality control procedures are based on QC of the timeseries (which are the exact same filtering methods as GSIM used) and an assessment of record lengths and missing data. However, you did not perform (or did not describe in the paper) any quality checks on the indices timeseries themselves. Were there outliers within regions, and if yes can you explain them? Can you determine the reliability of the indices based on the quality of the underlying

daily timeseries? What about abrupt shifts in the timeseries (i.e. from rating curve updates, instrumentation changes etc.)?

Response: Thanks for your questions. We have attached quality flags to every index value according to specific criteria. The flags represent the quality of index values (see L342-348). If you mean that we did not perform homogeneity tests, we think it is not necessary. Whether there are outliers within regions or whether there are abrupt shifts in the timeseries is beyond the scope of the manuscript. There are increasing studies focused on the non-stationarity of streamflow time series and attribution. The causes of inhomogeneity or non-stationarity could be manyfold and should be investigated with a more detailed case-by-case analysis (Tramblay et al., 2021). Our dataset is a good material for these studies. In terms of shifts caused by changes of measures, corresponding correction should have been done by data providers in the phase of compilation of database as only they know the details and how to perform a correction, which is out of the scope of our work. We could guarantee that the reliability of the indices values is determined by the quality of the underlying daily records, but the quality of the underlying daily records is only determined by the providers.

Great that you added flags. What I meant with 'quality checks on the indices' and 'outliers' was performing simple visual checks, by mapping the indices on a global map as I did in #9. This does not have to be an extensive analysis, but you can quickly identify some unrealistic data, if present.
True, the quality of the underlying records is first the responsibility of the provider. However, when creating a dataset such as this one, there is an opportunity as automated checks on for example non-stationarity can reveal issues that individual data providers were not able to catch. I agree that this is not absolutely necessary for publication, so this is your choice.

4. Indices based on baseflow estimates. The GSIM paper outlines certain issues relating to calculating indices based on baseflow, which made them decide to not include any. However, you provide recession indices without addressing any of the concerns outlined by GSIM.

Response: Thanks for your questions. We do not find relevant sentences in both Do et al. (2018) and Gudmundsson et al. (2018), but find sentences in Gudmundsson et al. (2018) as follows: "Note also that index selection was limited to those that can be computed without a base period, which excludes many; examples include "the number of days in a year, or season, for which daily values exceed a time-of-year-dependent threshold" (Zhang et al., 2005), drought deficit volumes (Loon and Anne, 2015; Tallaksen et al., 1997) and anomalies with respect to a climatological normal (McKee et al., 1993; Shukla and Wood, 2008). There are two reasons for excluding these indices". The term "base period" is not equivalent to baseflow.

You are right. My apologies for this oversight.

5. Overview and presentation of global indices. The paper misses a section giving an overall summary of global statistics for the indices generated. For example, a table providing the mean, max, and min of each index. Such a summary is also important as a quality check and can be used to compare the results to other studies that have provided global streamflow statistics.

Response: Thanks for your helpful advice. We have added a separate section (section 4 A comparative analysis) for comparative analysis on a global scale. Global trends in annual mean and percentiles of streamflow during 1970 to 2022 are mapped and compared with other studies' results (see L365-383 for details). Besides, we have provided global statistics of some indices named "Statistics.xlsx" in our dataset.

Great, I like this added comparative analysis.

6. Overall structure and presentation of the paper. Sections 3.2 and 4 are interesting but draw conclusions beyond the scope of this paper. For example, relating trends in streamflow to climatological drivers or land cover changes or other anthropogenic interference, without any robust analyses backing up these statements. Stick to a description of the dataset and a presentation of the data. Further on, the text contains many repetitions, use of casual language, grammar errors, and sentence structures that do not flow well. I recommend asking an external party to review your writing.

Response: Thanks for your advice. We have replaced the Section 4 with a comparative analysis mentioned in Reply #5. As to the Sections 3.2, we have retained and polished it since it gives an intuitive impression of our indices time series. The text and sentence structures have been improved.

OK I can see the improvements. Please avoid casual language however, such as "To make matters worse,…" (L248). Keep an objective tone.

7. I have accessed the .csv files only, as I do not use Matlab myself. The data is easy to access and to download, the overall description and citation information is clear, and the metadata is easy to find and well described. However, since this is a global-scale dataset which will attract researchers interested in large-scale comparative analyses, I would strongly recommend merging the 5548 separate time-series files into one csv and providing this file additional to the separate location-specific csv's. This way all information can be easily accessed using R or Python. Looking closely at a few of the individual files, they all start at the year 1806 and therefore contain much empty cells. I suggest removing the empty rows, merging the files and adding one column for station ID.

Response: Thanks for your recommendation. We have revised the dataset as you suggest. Indexspecific .csv files have been created with all stations in one .csv file. The empty rows in location-specific .csv have been removed. See our dataset for details.

Great!

8. The metadata contains information about catchment area. Does this refer to catchment area of the entire river reach or the contributing area upstream of the gauge location? Please specify (also in the paper).

Response: Catchment area refers to the contributing area upstream of the gauge location. We have specified it in the manuscript (see Table 5).

9. I mapped the MeanQ, Qmax, and Qmin and noticed a large region north-central Canada with no values, while they do contain values for other indices. This is a little suspicious to me. How can you calculate certain indices but not mean Q?

Response: Thanks for informing us. This is due to the difference of algorithms for different multi-year indices. Some multi-year indices, for example multi-year Qmean, were calculated by taking the average of corresponding yearly Qmean, while other multi-year indices, for example multiyear Q50th, were calculated by taking the median value of the whole daily time series. When there are lots of missing data in every year, all the yearly Qmean will be set to missing value, and so will the multi-year Qmean. In contrast, for multi-year Q50th, the missing ratio has no influence on the calculation based on the whole daily time series. We have revised the algorithms to keep these indices consistent.

OK. As mentioned also in #3, performing these kind of checks (i.e., visualizing on a global map) is not a lot of work, so make sure that you have done them also on other indices. In my opinion, it is not acceptable to have these relatively easily solvable algorithm errors in a published dataset.