Review of
"Simbi: historical hydro-meteorological time series and signatures for 24 catchments in Haiti" by Bathelemy et al. submitted to Earth Systems Science Data

1. Clarification

I put my focus mainly on the structure of the dataset. It is therefore advisable to have at least two more reviews for the text part (introduction, methodology, discussion).

2. General comment

I appreciate the work that was done by the authors. Haiti is a data scarce region and every initiative to homogenize, proof and publish available data should be encouraged. But there are high uncertainties in the dataset. It must be considered that the quality of a dataset has a very large influence on the results of working groups, which are using it! The applied methodology is not novel, but this isn´t in my opinion a must in a data journal. The text is easy to understand. Summarizing information for the individual gauges with graphical summary sheets (e.g. Figure 15) is a nice way for giving a quick overview. Nevertheless, there are some points that should be improved before publication:

3. Main comments

a. The period of the available data is far back. What happened with the gauges afterwards? How were the measurements done such a long time ago (e.g. devices and it`s accuracy)? Information about those questions would give a deeper insight into the used data basis.
b. Is there a schedule for digitizing the more recent observed (runoff) data? Are all (or at least the most) gauges still in operation? If so, it would be very nice to do the work in the context of this paper and update your results. I think that more recent results are more relevant than those from the period 1920 to 1940.
c. The uncertainty of the results is in some parts high (KGE values < 0,5 are often referred to as poor results). I would add a chapter "Uncertainties" in or after the results to address and describe all sources of uncertainties (e.g. long time ago, measurement method, data aggregation, model). I am aware that uncertainties are omnipresent with such a data basis, but from my point of view it would be helpful to list them clearly and centrally at one place in your manuscript.
d. In addition, I would make a very clear statement at the end and also in your abstract about the high uncertainties. Otherwise, there is the risk that further working groups do not sufficiently consider the high uncertainties.
e. The used data sources (e.g. land cover) and method for calculation the catchment attributes are not always clearly specified and listed. This is a lack of transparency.
f. Try to specify the data products used for calculating your catchment attributes (e.g. year of publication, domain, spatial resolution, update frequency, reference) at least in a list in the appendix.
g. The structure of your dataset isn´t friendly for machine-reading. Details are listed below.
h. Sharing the codes is helpful for reproducing your results and would make it easier for other groups to build upon your work.
i. Links should be listed in the references and not in the text, in order not to disrupt the reading flow.

4. Minor comments

They are just a first glimpse for the start, see clarification.

a. L45: The CAMELS datasets also provide indices and signatures.

b. L111: Can the digitization process be described (e.g. with an indication of accuracy)?
c. L177: Is the final error declared with an attribute?
d. L181: What Is indicated by the white, orange and blue dots? A legend would make it much easier to understand the figure.
e. L238: Why were the two individual datasets NOAA 20CR and BEST area-weighted averaged? Can a dataset be used without the other one?
f. L273: How was the aggregation done (e.g. area-weighting of the intersecting pixel or weighted all pixels equal not matter of the intersected individual area) and what was the output (e.g. mean, median). This information would be important.
g. L276: What is relevant to Haitian catchments and what not? I think the most readers do not have this information.
h. L279: Did you use the R-codes of Nans Addor (CAMELS-US) to calculate the hydrological signatures? If yes, it would be good to mention that.
i. L287: It seems that the two datasets for land cover are different data sources (other attributes). Can you give details about the differences (e.g. data genesis).
j. L288: There are global land use datasets (mostly derived by earth observation methods) for the most recent period.
k. L293: What is meant with “cut out”?
l. L367: What is the red line and the red point indicating in the figure?
m. L376: Try to be more precise, what is a critical percentage of missing data?
n. L392: If the relevant raingauge combination can cause such an error, have we to conclude that this amount of deviation is also possible at the other catchments? By the way: Can you declare the weights of the applied raingauge combination in a text file?
o. L423: There is partly a big difference of mean annual discharge between neighbouring gauges (> 600 mm/year). Can you explain that?
p. L470: This sentence fits better in the conclusion.
q. L484: How where the return periods calculated? Which statistical distribution?
r. L492: I would set the chapter Data availability after Conclusion.
s. L518: There are also outliers and not really a clear trend, see Figure 9.
t. L600: An additional column with the source dataset would enhance transparency in Table C1, C2 and C3.

5. Dataset

a. It is not user-friendly, that your dataset is compressed 2-fold. 1-fold (all files together) is enough.
b. I would name the compressed file Simbi and not “dataverse_files”.
c. The dataset is unnecessarily nested after extracting (e.g. dataverse_files\SIMBI\SIMBI_00_OBSERVED_DATA\00_OBSERVED_DATA\). Folder levels could be removed.
d. Storing the files as .csv instead .txt would enable an easy and fast opening with the software Excel. This is nice for a quick view.
e. The files are not friendly for machine-reading. I would remove the description in the header of your data files and create a new file with the belonging metadata. In this metadata file it is recommended to list the important information (e.g. Station-code, unit) directly after the name of the variable without blanks or special characters, e.g.:
   Station_code;P-001
   Unit;mm/d
   → Hint: Get rid of the blanks in the metadata, they are good for human reading, but not for machine reading.
f. It is tedious to read and separate lines like line 10 in Q_001___AMONT_DU_BASSIN___RIVIERE_GRISE (## Station code : 40501 ; Q-001 ; (DHSI ; URGEO). Try to get the DHSI station code (40501) with R, then you see why.

In general, try to make it as easy as possible to import data and the belonging metadata e.g. with the software R.

g. I would rename the files with the timeseries (e.g. instead Q_001___AMONT_DU_BASSIN___RIVIERE_GRISE → Q_1)
h. I would indicate the gaps in the time series not with a character (NA), simply do not enter anything. This ensures the data format numeric.
i. The ID of the gauge (code) should always be in the first column (this is not the case e.g. in the file location_and_topography.txt).
j. The number of decimal places is sometimes really too high (e.g. Percent_geologic_class.txt). This implies an accuracy which is not given.
k. The stream density includes a lot of information of a catchment (e.g. hint for carstic region). I would calculate and add this important catchment attribute. If there is no local dataset with the Haitian streams, take a global dataset like HydroATLAS (Linke et al., 2019, https://doi.org/10.1038/s41597-019-0300-6)
l. Have a look at Gudmundsson et al. (2018, https://doi.org/10.5194/essd-10-787-2018) regarding quality control flags (e.g. value does not change over more than x timesteps, outlier) for your timeseries.
m. Adding metadata to your shapefiles would be helpful, where the header and the unit of your attributes is declared.

6. Plots

a. The colour scales of your plots are not always intuitive (e.g. Figure 4 or 9). There is a nice article from Stoelzle and Stein (2021, https://doi.org/10.5194/hess-25-4549-2021) about visualization in hydrology. I recommend considering this guidance for your figures.
b. The font sizes in your plots vary sometimes greatly (e.g. Figure 7). It would be good to homogenize the sizes.
c. It is often more intuitive for the reader to extend the legend of a plot, than describing the attributes in the label of the figure (e.g. Figure 1).