Simbi: historical hydro-meteorological time series and signatures for 24 catchments in Haiti

Manuscript No. ESSD-2023-259

Ralph Bathelemy, Pierre Brigode, Vazken Andréassian, Charles Perrin, Vincent Moron, Cédric Gaucherel, Emmanuel Tric, and Dominique Boisson.

Reply to reviewer #1

We thank reviewer #1 for this detailed review and helpful comments. Please find below our replies to the reviewer's comments. We provided specific responses (in black) to the reviewer comments (in *italic and blue*).

1. **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:

   We thank the reviewer #1 for this positive feedback.

2. **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.*

      Streamflow measurements on Haitian rivers were conducted during the American occupation (1915-1934) under the supervision of USGS engineers. After the end of the American occupation in 1934, the number of active stations gradually decreased until 1940 when streamflow measurements ceased. Although some measurements were resumed in the 1960s and again in the 1980s, most of the data is now lost, and what remains is very fragmentary and only available for short periods. Consequently, the only reliable streamflow measurements are obtainable for the period 1920-1940.

      Graduated vertical rulers were installed on one riverbank and read two to three times daily. On average, each station collected 12 gauging measurements annually. Between 1925 and 1928, 11 of the 70 stations replaced the vertical rulers with automatic recorders; however, no information was available on the type of equipment installed. Hydrographic bulletins also reported on the quality of the rating curves (fair rating curve, very good rating curve for medium flows, fair rating curve for high flows, etc.).

      <u>This information will be provided in the next version of the manuscript. Additionally, metadata will include information about the instrumentation, whether it was manual or automatic, and the rating curve quality for each station.</u>

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.*

   After 1940, there is a scarcity of streamflow observation data available in Haiti. Good quality streamflow data can only be found for the period between 1920-1940, which has been presented in Simbi.

   <u>We will describe the evolution of hydrometry in Haiti since 1940 in the next version of the manuscript.</u>

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.*

   Thanks for this suggestion: <u>we will add a dedicated section devoted to the database uncertainties in the next version of the manuscript, to give comments on the various sources of uncertainties and their potential impacts. This should be helpful for the users of the datasets.</u>

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.*

   <u>Uncertainties will be clearly stated in the conclusion of the manuscript and in the abstract in the next version of the manuscript, to warn working groups of the possible pitfalls associated with the use of this dataset.</u>

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.*

   <u>The next version of the manuscript will provide a more detailed description of the data products used, including land cover, geological, and aquifer type data. Additionally, Table C2 will include the mathematical formulations and references for each attribute used.</u>

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.*

   The catchment-scale hydro-climatic data produced in this article and the land-use, geological, aquifer type and topographical data were used to calculate the attributes.

   <u>The next version of the manuscript will list the information of the data products used to calculate catchment attributes (e.g. year of publication, domain, spatial resolution, update frequency, and reference).</u>

g. *The structure of your dataset isn´t friendly for machine-reading. Details are listed below.*

   The structure of the dataset has been modified for machine-reading (see the dataset section below).

h. *Sharing the codes is helpful for reproducing your results and would make it easier for other groups to build upon your work.*

R codes reading Simbi on all catchments and performing rainfall-runoff model calibration and simulation over the entire period will be provided with the next version of the manuscript.

i. *Links should be listed in the references and not in the text, in order not to disrupt the reading flow.*

The links will be removed from the text, except for the link in the abstract because the dataset link must be included in the abstract according to the ESSD manuscript preparation guidelines.

3. **Minor comments**

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

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

The sentence will be replaced by: *"While the CAMELS databases provide time series, indices and hydroclimatic signatures of catchments, other databases provide only indices and hydroclimatic signatures of catchments, such as the African Database of Hydrometric Indices (ADHI; Tramblay et al., 2021)".*

b. *L111: Can the digitization process be described (e.g. with an indication of accuracy)?*

The digitization process is described in Appendix A. However, it will be described in greater detail in the next version of the manuscript.

c. *L177: Is the final error declared with an attribute?*

The final error will be declared with an attribute in the next version of the manuscript.

d. *L181: What Is indicated by the white, orange and blue dots? A legend would make it much easier to understand the figure.*

The white, orange and blue dots indicated all raingauges with monthly data for the period 1920–1940. Raingauge stations with air temperature data are shown in orange. Raingauges considered relevant for hydrological modeling are shown in blue (see section ***Erreur ! Source du renvoi introuvable.***). Other raingauges are shown in white. A legend will be added in the next version of the manuscript.

e. *L238: Why were the two individual datasets NOAA 20CR and BEST area-weighted averaged? Can a dataset be used without the other one?*

Both air temperature databases were used independently. The wording of the previous sentence was unclear, so it will be revised in the upcoming version of the manuscript:

Catchment air temperature series were computed on a daily time step for two temperature databases (NOAA 20CR and BEST) by taking the weighted average of pixels in the respective database (NOAA 20CR or BEST).

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.*

Catchment scale data are computed as a weighted average. The weights are proportional to the area of the pixel overlapping the catchment. This chapter will be described in greater detail in the next version of the manuscript.

g. *L276: What is relevant to Haitian catchments and what not? I think the most readers do not have this information.*

Thanks for this suggestion. This sentence will be reworded:

A set of attributes that describes a broad range of low, moderate, and extreme precipitation and streamflow characteristics were chosen to characterize the hydrological regime, based on available data.

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.*

   We did not use the R codes of Nans Addor to calculate hydrological signatures.

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).*

   Thanks for this suggestion. Details of the differences between the two land cover datasets will be provided in the next version of the manuscript.

j. *L288: There are global land use datasets (mostly derived by earth observation methods) for the most recent period.*

   The existence of these alternative datasets will be mentioned in the next version of the manuscript.

k. *L293: What is meant with "cut out"?*

   The expression "cut out" will replaced by "cropped".

l. *L367: What is the red line and the red point indicating in the figure?*

   The red line represents the optimal ratio (r=1), while the red dot represents the mean value of the distribution. The figure caption will be modified:

   Ratio of the GR2M calibrated parameters X1 and X2 over the two subperiods for the reference and relevant raingauge combinations. The red line represents the optimal ratio (r=1), while the red dot represents the mean value of the distribution.

m. *L376: Try to be more precise, what is a critical percentage of missing data?*

   If a station has less than 10 years of data, it will be deemed to have a critical percentage of missing data (The 10-year threshold was set arbitrarily). This paragraph will give more detail in the next version of the manuscript.

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?*

   If the relevant rain gauges are wet or dry, a greater or lesser deviation may be observed. The weights of the rain gauge combination applied will be presented in a dedicated file.

o. *L423: There is partly a big difference of mean annual discharge between neighbouring gauges (> 600 mm/year). Can you explain that?*

   The Q-008 watershed has a significantly higher mean annual streamflow than the three neighboring watersheds (Q-010, Q-068 and Q-029). As indicated in line 387, over 90% of the Q-008 watershed is located on a limestone geological formation, and there is also an impact from karst aquifers within the watershed. On this basis, it is likely that an influx of water from neighboring catchments is responsible for such a high mean annual streamflow. However, no studies have been carried out to verify or challenge this hypothesis. This point will be highlighted in the next manuscript version.

p. *L470: This sentence fits better in the conclusion.*

   Thanks for this suggestion. This sentence will be moved in the conclusion.

q. *L484: How where the return periods calculated? Which statistical distribution?*

   The Generalized Extreme Value distribution and the distribution of annual values (precipitation, PE, air temperature or streamflow) were used to estimate values for multiple return periods. This will be clarified in the next manuscript version:

   Simulated streamflows underestimate maximum annual flows with a return period of less than 10 years and overestimate flows beyond 10 years. The generalized extreme value (Beirlant et al., 2004; Coles, 2001; Jenkinson, 1955) and the distribution of annual *values (precipitation, PE, air temperature or streamflow) were used to estimate values for multiple return periods.*

r. *L492: I would set the chapter Data availability after Conclusion.*

The data availability section must be placed before the conclusion section according to the ESSD manuscript preparation guidelines.

s. *L518: There are also outliers and not really a clear trend, see Figure 9.*

Thanks for this comment. Indeed, the central part of Haiti is associated with low streamflow values and the southwest with high streamflow values. However, no clear trend was observed in the north. This analysis will be changed in the next manuscript version:

The central part of Haiti is associated with relatively low streamflow and high drought coefficients. The southwest is associated with relatively high streamflow. In fact, large floods are more frequent in these areas (Terrier et al., 2017). No clear trend was observed in the north.

t. *L600: An additional column with the source dataset would enhance transparency in Table C1, C2 and C3.*

An additional column with the source dataset will be added in the next version of the manuscript.

4. Dataset

a. *It is not user-friendly, that your dataset is compressed 2-fold. 1-fold (all files together) is enough.*

Thank you for your comment. Unfortunately, this two-time compression is due to the data warehouse that adds extra folders, and we are not able to change that.

b. *I would name the compressed file Simbi and not "dataverse_files".*

The folder "dataverse_files" is added by the data warehouse and is not reliant on us.

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.*

The first three levels (dataverse_files\SIMBI\SIMBI_00_OBSERVED_DATA) are automatically included by the data warehouse and therefore cannot be removed. These directories are intended for use by other databases also hosted within this warehouse, such as the ADHI database (https://doi.org/10.23708/LXGXQ9).

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.*

The files are now saved in csv format.

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.*

The headers of the data files have been deleted and new files have been created to store the metadata. The metadata files include:

i. a file containing information on rain gauge stations,
ii. a file containing information on streamflow stations,
iii. a file containing monthly time-step modeling information, and
iv. a file containing daily time-step modeling information.

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.*

   Metadata is now stored in separate files in a tabular format, allowing for easy importation.

g. *I would rename the files with the timeseries (e.g. instead Q_001___AMONT_DU_BASSIN___RIVIERE_GRISE → Q_1)*

   Files have been renamed using station codes to facilitate identification (e.g. Q_001 instead of Q_001___AMONT_DU_BASSIN___RIVIERE_GRISE).

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.*

   In certain files, one column contains observed streamflows (including gaps) while another column contains simulated streamflows (without gaps). This discrepancy complicates the removal of rows with missing observed data. To maintain numeric consistency, the NA character is now substituted with the value of -9999.

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).*

   Thanks for this comment. Station IDs are now in the first column.

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.*

   Thanks for this comment. The decimal places are now limited to 2.

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)*

   Thanks for this suggestion. Stream density will be included as an attribute in the next version of the manuscript by utilizing dataset from CNIGS (National Center for Geospatial Information in Haiti).

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.*

   Thanks for this suggestion. In addition to the criteria outlined in section 3.1.1 for selecting streamflow series, the three indicators proposed by Gudmundsson *et al.* (2018) for quality control of time series have been added in Simbi:

   i. days for which Q<0, where Q denotes a daily streamflow value,
   ii. daily values with more than 10 consecutive equal values larger than zero, and
   iii. outlier detection.

m. *Adding metadata to your shapefiles would be helpful, where the header and the unit of your attributes is declared.*

   Metadata (name, code, altitude, and area) and their units will be added to the shapefiles.

5. 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.

Thank you for this comment. The next version of the manuscript will include modified color scales for the figures.

b. *The font sizes in your plots vary sometimes greatly (e.g. Figure 7). It would be good to homogenize the sizes.*

The font sizes of the plots will be homogenized in the next manuscript version.

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).*

Figure attributes will be described in the figure legend in the next manuscript version.