Review of “HARBOR - Harmonized Attributes for River Basins in One Repo: Collated River Basin Data from Multiple Collections with a Software Toolkit”

Paper ID https://doi.org/10.5194/egusphere-2025-5786

General comments

The paper presents a novel framework to aggregate and harmonize catchment attributes from multiple US datasets. The historical fragmentation of basin data has been a significant bottleneck in large-sample hydrology, and I think this contribution is timely as it provides a valuable roadmap for researchers navigating multi-source data.

The core value of HARBOR lies in the data collection and harmonization framework, and I suggest concentrating the efforts on strengthening these aspects. The work is potentially very useful, but I think it is more suited for another journal, e.g. Earth System Science Data (ESSD), as it focuses on furthering the reuse of high-quality data (which is in the ESSD aims and scope) rather than on the improvements of hydrological process understanding.

Additionally, I believe the manuscript, in its current form, requires further structural clarity and methodological depth. Specifically:

- Scope and focus: The current Abstract does not sufficiently define the scope of the framework. The term 'attributes' is too vague; it is essential to specify which data categories are included (e.g., physiographic, climatic, land-use, etc.) to immediately inform the reader.
- Data Presentation and Accessibility: The current presentation of data and results is difficult to navigate. To make the framework truly accessible, the authors should avoid long walls of text and overly dense tables with extended fields. I suggest a more “scannable” structure, using concise summaries and well-designed tables to improve readability for the final user.
- Methodological Robustness: The section introducing new analyses (such as the SWB extension) appears less robust compared to the rest of the paper. As these new findings rely on simplified assumptions and lack proper validation, I recommend removing this part to keep the focus entirely on the core contribution: the data collection and aggregation framework.

Therefore, I encourage a resubmission to another journal, such as Earth System Science Data (ESSD), as its focus on data and tools aligns much better with the strengths of this contribution.

In the following sections, I listed specific suggestions that I hope will be useful in improving the manuscript.

Specific comments

On the repository:

- I suggest unifying all project materials under a single GitHub repository named after the framework (HARBOR). The current split between repositories with different names is confusing; a consolidated repository would better represent the project's identity and simplify the maintenance of the 'final' version of both data and code.
- While .TSV is technically accurate for tab-separated files, the .CSV extension is far more ubiquitous and recognized by most operating systems and users as the standard for tabular data. I suggest renaming these files to .CSV to improve immediate recognition, while clearly specifying in the documentation or the file header that the delimiter used is a tab.
- The presence of multiple symbols for missing values (empty, -9999, '.', '-') can lead to errors during data processing. Please provide an additional version of the final collated dataset using a unique and consistent null-value indicator to improve interoperability and prevent misinterpretation by end-users.
- The use of the Python 'pickle' format for data distribution is discouraged due to significant security vulnerabilities (https://docs.python.org/3/library/pickle.html#data-stream-format) and lack of cross-language compatibility. Please transition to a more robust, interoperable format such as JSON.
- Please provide also an open-document format of all .xlsx files. This will guarantee better long-term data preservation and facilitate access for users who do not use the Microsoft Office suite.
- Please replace all .webloc files with open-format text files. For a project aiming for broad community adoption, it is important to avoid proprietary metadata formats and stick to universal open standards for all repository contents.

Line comments:

- 70-78: When outlining the structure of the paper, please refer to sections by their numbers (e.g., 'Chapter 2') in addition to their titles. This will significantly improve internal navigation for the reader.
- 80 (Chapter 2): The review of existing data collections would be more effective if presented in chronological order. A chronological flow provides a better logical progression and avoids confusing forward-references (e.g., mentioning MOPEX at line 87 before it has been formally introduced).

  To improve visual clarity, I suggest using an indented bulleted list for these datasets; the current paragraph format makes it difficult to distinguish between section headers and descriptions. Additionally, the paper would benefit greatly from a summary table (listing name, basin count, reference status, and available time series) and a timeline figure to provide a quick overview of the state of the art.
- 291: Please explicitly list which specific datasets were merged to create the final collated file. A brief summary or a small flowchart would ensure the provenance of the final attributes is transparent.
- 301: Similar to Chapter 2, the list of hydrologic signatures would be much more accessible if formatted with bullet points. This will help the reader quickly identify the specific metrics included in the framework.
- 327 and 360: Tables 2 and 3 could be significantly improved for better readability. I suggest splitting the long, concatenated class names into individual columns (e.g., Climate,

Topography, Soil Permeability, and Bedrock Permeability). This multi-column approach allows for easier comparison and filtering of the classification criteria.

- 388-398: Methodology for SWB class extension:

  The proposed extension of the Seasonal Water Balance classification relies on simplified assumptions and datasets whose origin is not fully documented. Without a robust validation against established benchmarks, the reliability of this extension remains unverified. As stated in the general comments, my primary recommendation is to omit this part and focus on the data harmonization. However, specific comments are listed below.

  a) Nomenclature Confusion: It is unclear whether "GAGES-II CONUS" refers to the "SB3" collection (line 208) or the combined set “USGS GAGES-II” (line 195), filtered to the CONUS area. Please specify how the number of basins (9067) is obtained. Consistent naming is essential for reproducibility.

  b) Geometric vs. Hydrologic Extension: Expanding the 10 SWB classes by simply stretching their boundaries to fill the index space (Fig. 8) lacks hydrological justification. Extrapolating a classification originally built on ~300 MOPEX basins to over 9,000 basins without discussing the increased complexity or the mixed provenance of data is a major methodological flaw.

- 415-439: Proxy-based estimation of seasonality

  a) The authors estimate precipitation seasonality by inverting a formula from Woods (2009). However, this relationship was calibrated on a very limited sample (6 stations); applying it to 9,000 basins across diverse CONUS climates without sensitivity analysis or validation is statistically unsound.

  b) "800 m PRISM data" is mentioned without citation or documentation. The authors could clarify the data source. Additionally, the snow-rain threshold lacks supporting evidence and contains typographical errors (1 °C).

- 453-464: Qualitative validation of the SWB extension.

  a) Validating class A1 based on anecdotal conifer distribution is insufficient. The locations mentioned are not labeled in the figures, making it difficult for an international audience to verify the claims. The authors should use formal vegetation maps for a rigorous spatial comparison, rather than relying on qualitative geographic descriptions.

Comments on the figures:

- Figures placement: For the sake of readability, please ensure all figures are embedded within the manuscript body, positioned as close as possible to the relevant text.
- Figure 1: It is unclear if the size of the square markers reflects the density or number of basins in each region. If not, I suggest scaling the markers proportionally to the basin count or adding a secondary "bubble map" to visually represent the dataset's spatial density.
- Figure 5: To make this figure self-explanatory, please include a legend or a small inset table summarizing the class names. This will save the reader from having to scroll back to the main text to interpret the symbology.
- Figure 10: I recommend removing the outer bounding box and the geographic coordinates. Cleaning these elements will reduce visual clutter and allow the focus to remain entirely on the spatial data presented.

Line typos:

- 172: Please check the sentence
- 601: Please check the sentence