Authors' Response (Final)

2024-07-02 - Daniel Kovacek & Steven Weijs

Authors' thanks: We would like to express our sincere gratitude for the time and effort of the editors and anonymous referees in reviewing and providing feedback on our work. The points raised by all highlight important clarifications, and the quality of the revised manuscript is significantly improved as a result.

Final responses to all referee comments to the draft manuscript are detailed below. The information is organized in the following order: i) referee comment, ii) author response, and iii) line numbers and/or sections identifying related manuscript changes. Please note that page and line numbers referring to manuscript edits correspond to the **revised manuscript**.

Responses to RC1 comments

RC1 Comment: *A definition of the basin considered in this study is needed. Basin is a term that is interchangeable with catchment and watershed, but it typically refers to the entire drainage area of a river. In this article, 'basin' represents the local watershed of each river-reach. The term 'sub-catchment' or 'sub-basin' is more appropriate here.*

Author's Response: While we agree that many of the basins considered in our dataset could be classified as sub-basins or sub-sub-basins, we use the term basin in a wider sense of the definition. This is in line with literature about ungauged basins. For example, the usage of "basin" in *"A decade of Prediction in Ungauged Basins (PUB)--a review"* (Hrachowitz et al., 2013) does not seem to refer exclusively to the entire drainage area of a river. We agree there should be an explicit definition of our use of the term "basin", and it has been added at the start of Section 1.2.

Corresponding Manuscript Edits:

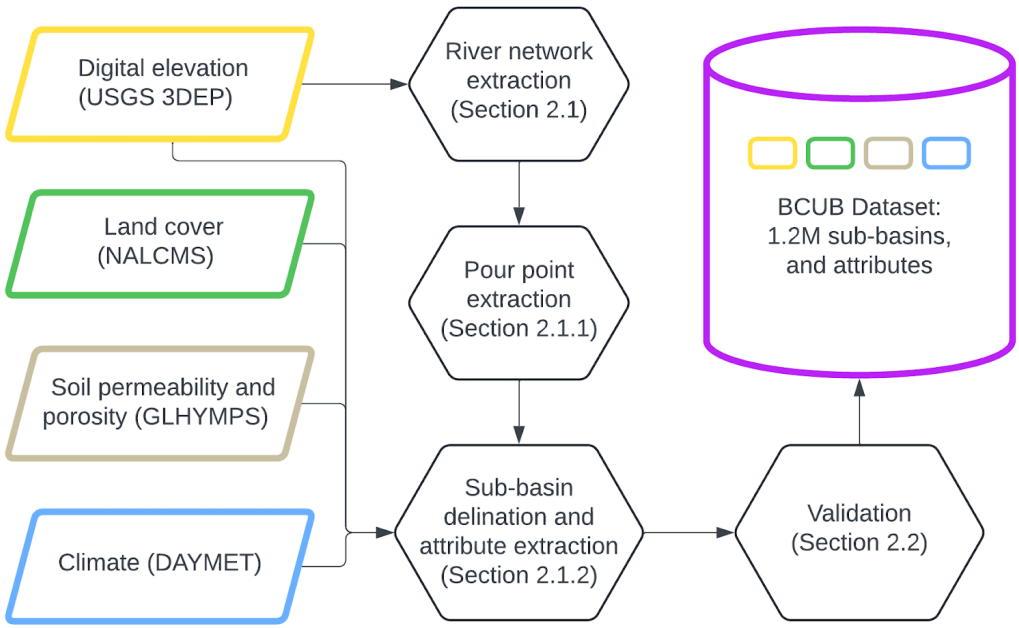
- An explicit definition of our usage of the term "basin" as "the local watershed of any confluence or outlet in a stream network." has been added at the start of Section 1.3 (line 85).
- All references to "basin" have been reviewed and changed to "sub-basin", "catchment", or "watershed" where appropriate to the specific context.

***RC1 Comment**: To understand the process more easily, a flowchart showing different steps of BCUB database development in the methodology section would be helpful.*

Author's Response: We agree a diagram will provide a useful overview of the full process.

Corresponding Manuscript Edits:

- The diagram below has been added as Figure 2 to the manuscript (top of page 5) to represent the dataset development process:



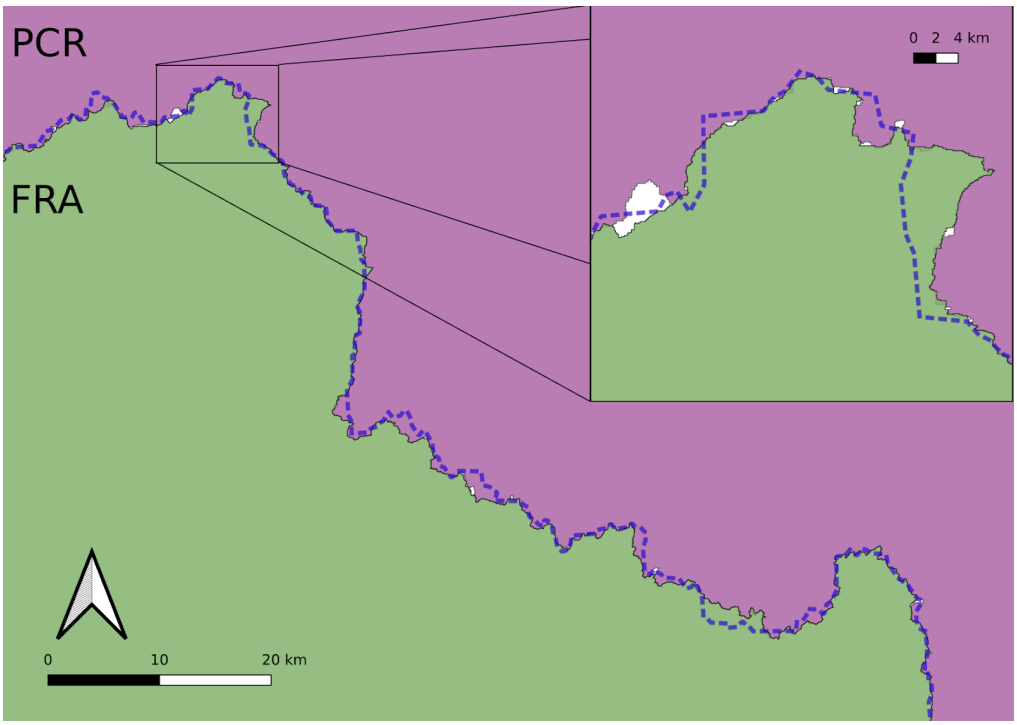
***RC1 Comment**:The reason for using the HydroBASINS watersheds (level-5 and 6) to subdivide the study region is understandable. However, the underlying hydrography data in the HydroBASINS and BCUB databases are different. So, there is a chance of missing a part of the sub-catchments located near the regional boundary in the BCUB database. For example, a part of the sub-catchment of the PCR region, located near the boundary between PCR and FRA, may overshoot to the FRA region due to hydrographic data inconsistencies. How was this issue addressed during the development of this database?*

Author's Response: Thank you for raising this important point. While the region polygons assembled from HydroBASINS are a helpful tool for organizing the data processing pipeline, indeed their use yields different bounds whose effect on sub-basin delineation is in the order of the size of the smallest sub-basins in the BCUB dataset.

The Caravan dataset (Kratzert et al., 2023) clearly describes the issue with aggregating attributes from catchment polygons that do not align with the HydroBASINS dataset. By independently deriving the region bounds from a single continuous DEM source (USGS 3DEP, 30m grid), we

avoid the problem of misalignment with HydroBASINS polygons, however it does not solve the problem of uncertainty in region bounds defined independently of HydroBASINS.

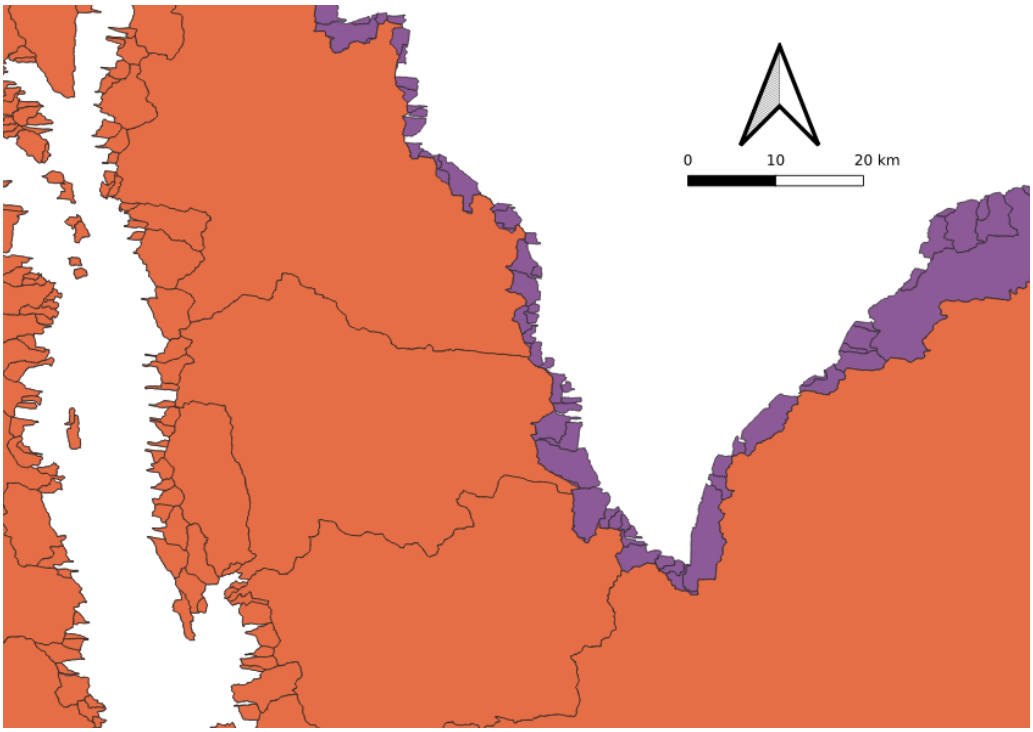
Below is an outline of the process we use to independently redefine sub-region polygons from the DEM and quantify uncertainty in region bounds in the BCUB dataset.



The edge detail inset in the figure above shows an example segment of region boundaries aggregated from HydroBASINS (blue dashed line) compared to independently derived region bounds. The purple (Peace, PCR) and green (Fraser, FRA) coloured areas represent the region boundaries derived independently using the USGS 3DEP DEM (30m grid resolution), referred to here as the BCUB region boundaries. White areas are gaps that remain following the iterative boundary definition process described below. We define boundary deviations as polygons representing i) gaps between region bounds where the DEM resolution does not resolve which direction the small area drains, and ii) boundary overlaps when delineating from pour points in distinct basins with shared boundaries.

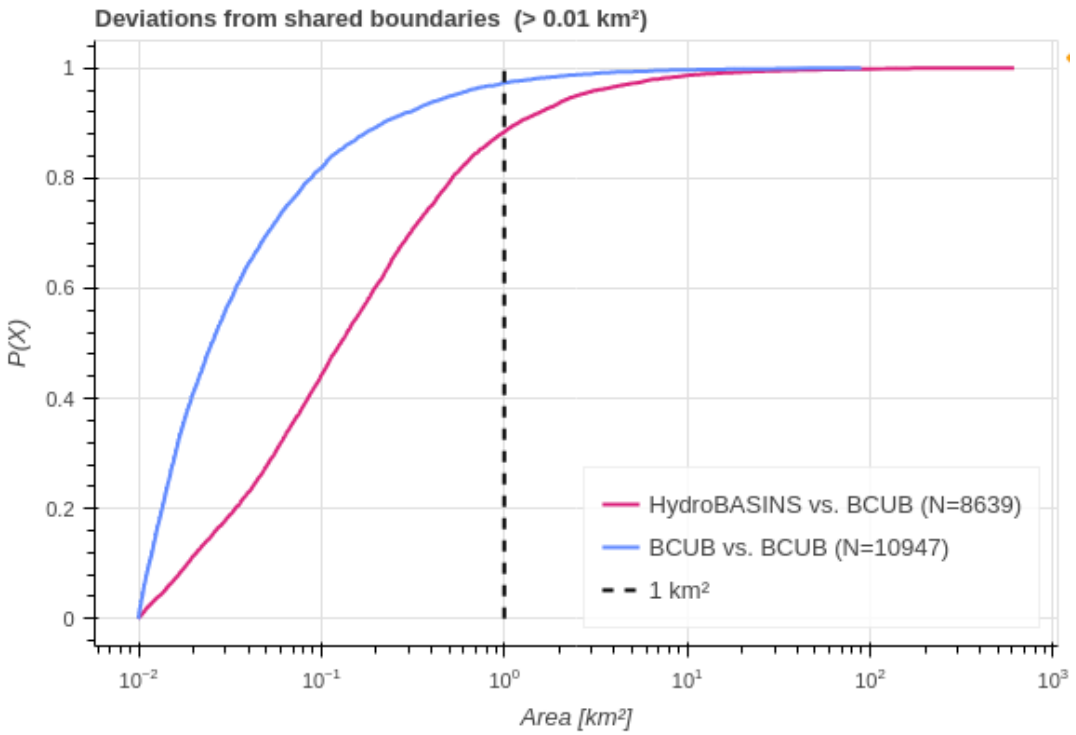
The process begins by applying a (5km) buffer to the region boundaries aggregated from level 5 and 6 HydroBASINS polygons, and using these buffered polygons as clipping masks on the DEM. The purpose of this step is to avoid restricting the catchment boundary delineation by the clipping mask. The covering set of polygons (catchments) are then delineated from the clipped DEM for each region, and the exterior edges (of the union of intersecting geometries) are checked to verify that they do not touch the edge of the buffered region polygon. Where the edges intersect, we manually expand the buffer (DEM clipping mask) in QGIS and re-derive the covering set of catchments until the buffer is sufficient, i.e. the covering set of basins does not touch the edge of

the clipping mask. The use of a buffer causes small catchments to be delineated which drain to basins in adjacent regions, and these are excluded by identifying that they are completely contained by the clipping mask of the neighbouring region. The figure below illustrates the excluded vestigial edge sub-basins (purple) and the remaining covering basin set (orange).



Delineating region boundaries independently from the covering set of basins does not yield perfectly shared boundaries, but these deviations are substantially smaller than those resulting from aggregating the HydroBASINS levels 5 and 6 polygons. The distribution of the size of deviations from shared sub-region boundaries are shown in the figure below. The red series represents deviations between the BCUB region bounds and HydroBASINS-derived bounds (median area of 0.13 km^2), while the blue series represents deviations (overlaps and gaps) within the BCUB sub-region boundaries (median area = 0.03 km^2). Polygons smaller than 0.01 km^2, or 1% of the smallest sub-basin in the BCUB dataset were neglected.

We will incorporate a geometry flag attribute in the BCUB dataset for any sub-basin that intersects or touches at least one boundary deviation, and will include a decimal value to represent the total deviation area as a percentage of the sub-basin area. Where two different sub-basins claim the same area, either bordering sub-basin may overestimate the catchment area (indicated by a positive % value). Where an area is not claimed by any basin but is not necessarily endorheic, either bordering sub-basin may be underestimating the catchment area (indicated by a negative % value). The percentage represents the maximum expected percentage error from the uncertain boundary. The purpose of including these quantities is to communicate (some part of) the uncertainty in defining region bounds where the size of the uncertain area exceeds 1% of any sub-basin area. We will update the region boundaries in the data repository, and we will



additionally provide the set of polygons representing boundary deviations as a .geojson file to facilitate corrections given updated information resolving these disagreements.

We additionally point out that a precise coastline definition (or ocean masking) at the resolution of the input DEM is important for the river network processing computation, otherwise vestigial river segments occur in the ocean parallel to coastlines where the HydroBASINS polygons extend over ocean surface. We crop the coastline using the NALCMS land cover data ocean pixels – the land cover data are well suited to the input DEM since the both products are provided in the same grid resolution.

Finally, these region boundary updates will require revising the BCUB dataset. We will reprocess all affected sub-basins and update the dataset with the above additional information, namely the catchment delineation flag and the percent area represented by uncertain region boundaries. The additional detail provided here will appear in some form in the manuscript. The code used to derive the region boundary deviations will be provided along with the existing validation code in the open-source code repository. We believe these revisions will result in a more transparent and higher quality dataset, and we appreciate the reviewer raising this important detail.

Corresponding Manuscript Edits:

- Section 2.2.1 (page 10-11) has been added to describe the treatment of uncertain sub-region boundaries.
- Figure 5 has been added at the top of page 11 to illustrate the problem of uncertain region bounds,

- Figure 6 has been added at the top of page 11 to quantify the effect of the treatment described in section 2.2.1 on reducing the size of boundary deviations.
- Three additional columns have been added to the dataset to (1) flag where uncertain boundaries border with catchments in the BCUB and to quantify the uncertain area (2-gap, 3-overlap) as a percentage of the catchment area. Definitions of these attributes have been added to subsection 2.1.2 at lines 161 to 166.
- Table 3 has been updated to reflect the additional metadata attributes (top of page 22).

RC1 Comment: *It is sometimes difficult to follow the article due to inconsistencies in the statements. For example, the line 76 in the motivation section, "The accuracy of stream network delineation improves with increasing DEM resolution." The transition from the previous lines to this one is not smooth.*

Author's Response:
We agree.

Corresponding manuscript edits:

- The point about accuracy of stream networks was moved to Section 2.2 (line 190) where it is more relevant.

RC1 Comment: *Another example of inconsistency is in line 134, where the delineation of the stream network is discussed after the description of the pour point selection process from the stream network. It would be more appropriate to discuss the stream network delineation process before selecting pour points.*

Author's Response:
Agreed. The order of stream network extraction and pour point selection have been adjusted accordingly to improve the consistency overall narrative and sequencing of arguments.

Corresponding manuscript edits:

- A point-by-point ordered summary of the data collection and processing has been added to the introduction of section 2.1 (page 5, lines 112-131). The more detailed information about points 4-6 (lines 123-130) have been reordered in subsections 2.1.1 (line 132), 2.1.2 (line 153), and 2.1.3 (line 167) to correspond with the sequence in which they are introduced.
- Small edits to have been made throughout the manuscript to improve the overall grammar and organization with deliberate care to preserve the content, meaning, and interpretation of the arguments.

RC1 Comment: *Line 103: Please provide the minimum drainage area threshold used to delineate the stream network from USGS 3DEP*

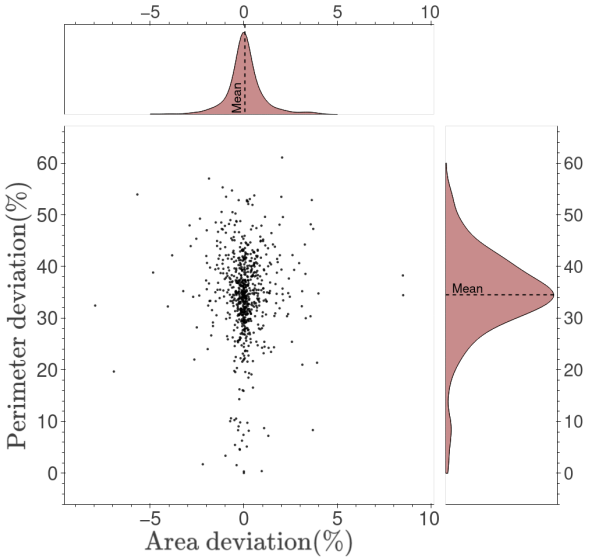
Author's Response: The minimum drainage area threshold used is 1 km^2 which corresponds to the smallest sub-basin included in the HYSETS dataset (Arsenault et al. 2020) and to the smallest monitored basin in the British Columbia streamflow monitoring network. This reference is made explicit in the text, but your note identifies where (we agree) it should be placed earlier in the text.

Corresponding manuscript edits:

- We have moved the explicit reference to minimum drainage area threshold earlier in the text as recommended to line 109.

RC1 Comment: *Figure 7: This is a nice figure to show the impact of using DEM with different resolutions. The plot with colored density would be more helpful to understand the figure.*

Author's Response:
Figure 7 (Figure 10 in the revised manuscript) has been modified (see below) to show the probability densities in both x and y, which we believe adds clarity to the meaning of the figure. We tried a 2D (kernel) density plot to unsatisfactory effect due to either requiring a less interpretable colour mapping or x and y units. We believe the addition of probability densities of x and y are a reasonable compromise to effectively communicate the point that when increasing the input DEM resolution, the mean change in area is near zero while the corresponding change in perimeter is substantially greater than zero. We also add here that the coefficient of determination (R^2) between x and y (area and perimeter deviation from baseline) is zero.

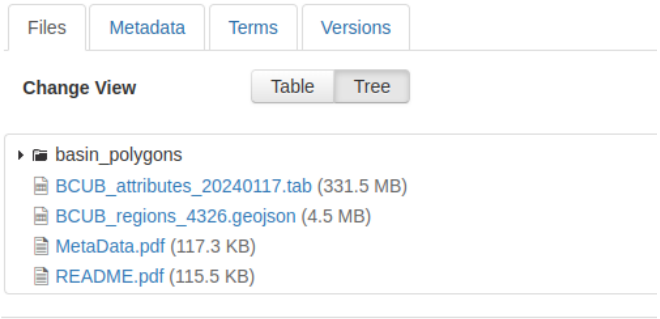


Corresponding manuscript edits:

- Figure 7 in the draft manuscript is now Figure 10 (shown above), located at the top of page 15.

RC1 Comment: *When using QGIS version 3.28 to open the dataset, it displays the pour point location instead of the sub-basin polygon. Has the delineated sub-basin geometry been excluded from the database?*

Author's Response: The tabular file (BCUB_attributes_20240117.tab) contains the x,y coordinates of the pour point (ppt) and basin centroid ('centroid_x', 'centroid_y', 'ppt_lon_m_3005', 'ppt_lat_m_3005') while due to the very large file sizes, the polygon geometries are provided separately in the Parquet file format saved under the "basin_polygons" folder in the data repository:



To avoid issues with limited memory (<64GB), the default geometry in the parquet files is set to the pour point to use considerably less memory than the polygon geometries. We recommend filtering geometries based on specific questions before loading polygons for visualization.

Parquet is supported by GDAL as of version 3.5, so QGIS must be compiled with GDAL >= 3.5 which is not default in some environments.

Please see the following for information about versions and compatibility:
https://gis.stackexchange.com/questions/430973/importing-geoparquet-file-in-qgis

Reading/writing Parquet in R:
https://arrow.apache.org/docs/r/reference/read_parquet.html

Reading/writing Parquet in Python:
https://arrow.apache.org/docs/python/parquet.html

Parquet is also implemented in Julia, MATLAB, Rust, Go, Java, C++, and others:
https://arrow.apache.org/docs/

Corresponding manuscript edits:

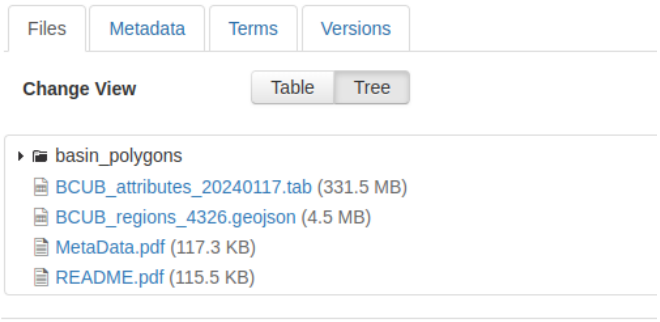
- Suggested resources for working with the data have been added in lines 312-317.
- A reference pointing to the notebook tutorial demonstrating import and use of data from the Parquet format has been added at line 331.

Responses to RC2 comments

RC2 Comment: *It would be nice if authors can convince us the necessity to have 1.2 million basins (sub-basins or sub-catchment), it's very difficult to find the one you are interested and very difficult to use all of them in a regional scale.*

Author's Response: The smallest monitored sub-basin operated by the BC Hydrometric Service (and likewise in the HYSETS dataset) is 1 km^2. In developing the BCUB dataset, we aimed to cover the range of basin sizes described by the set of monitored watersheds in BC since these are widely used in research and practice. The reason for this is that we want to comprehensively characterize the ungauged space, i.e. the set of ungauged basins above the size threshold. This will serve further research in what we might be missing with the current monitoring network. We will aim to highlight this goal somewhat more in the introduction.

We agree that the large number of sub-basins does create challenges for working with the dataset as a whole. We provide an example of working with (parts of) the data in a web-based Jupyter Notebook: (https://dankovacek.github.io/bcub_demo/notebooks/7_Dataset_plot_example.html) In an effort to support different use cases, we provide the data in two formats: i) a smaller and widely used (csv) format containing all sub-basin attributes with x,y coordinates of the sub-basin pour point and polygon centroids, and ii) the much larger basin polygon files in (geospatial) Parquet format (saved under basin_polygons in the data repository).



The Parquet format is supported by GDAL as of version 3.5, so QGIS must be compiled with GDAL >= 3.5 which is not default in some environments.

Please see the following for information about versions and compatibility:
https://gis.stackexchange.com/questions/430973/importing-geoparquet-file-in-qgis

Reading/writing Parquet in R:
https://arrow.apache.org/docs/r/reference/read_parquet.html

Reading/writing Parquet in Python:

https://arrow.apache.org/docs/python/parquet.html

Parquet is also implemented in Julia, MATLAB, Rust, Go, Java, C++, and others: https://arrow.apache.org/docs/

Corresponding manuscript edits:

- The aim of covering the range of catchment areas represented in large sample hydrology is described in lines 45-49. The number of basins in the BCUB is a function of aiming to cover the range of spatial scales described by related datasets.

RC2 Comment: *If possible, authors can introduce some specific implementations of these large-sample basins.*

Figures 9 and 10 in the manuscript give examples of specific questions that can be asked of this dataset. The ability to derive customized samples of basins by a wide range of characteristics may support experimental design, for example in temperature monitoring for quantifying the effect of land cover change on stream temperature. We are currently using the BCUB dataset as an input for a streamflow monitoring network optimization study.

Corresponding manuscript edits:

- Suggested resources for working with the data are provided in lines 312-316.
- A link to and description of the notebook (tutorial) demonstrating import and use of data from the Parquet format has been added in lines 329-332.

RC2 Comment: *It is also very difficult to find real observation of river discharge to support further analysis.*

Author response: The HYSETS dataset (Arsenault, 2019) provides streamflow observation at a large set of monitored locations along with their catchment polygons and attributes, whereas the BCUB dataset defines and describes attributes for sub-basins at all river network confluences and does not contain streamflow since the vast majority of confluences are ungauged. These two datasets can be combined to extrapolate information from monitored to unmonitored catchments. It should be noted that some work is required to map locations between HYSETS and BCUB datasets, and many historical monitoring locations no longer exist and their location coordinates were recorded with varying degrees of precision.

RC2 Comment: *Please check the unit of precipitation, e.g., 2028mm/day for gauge 1269663 must be wrong?*

Author response: Thank you for catching the typo, the precipitation index represents a mean annual value, which is derived from daily total precipitation from the DAYMET dataset. We have updated Table 2 to read "Mean Annual Precipitation" with the corresponding units [mm/year].

Corresponding manuscript edits:

- Table 2 (page 21) has been updated with the correct units (Annual precipitation, mm/year)

RC2 Comment: *What is the difference in the number of detected sub-basins when using the two spatial resolution?*

Author response: The number of sub-basins is a function of the minimum area threshold assumed in the stream network extraction process, which we set at 1 km^2 to match the smallest catchment in the streamflow monitoring network. Our hypothesis is that the number of basins is not expected to change significantly as a result of a change in the input DEM resolution. The key factors to consider are:

1. The processing of DEM and flow direction raster data to define the stream network assumes a minimum area threshold.
2. The number of raster cells (pixels) representing the smallest sub-basin (1 km^2) increases by roughly an order of magnitude between the EarthENV (90m at the equator, ~10^2 pixels) and the USGS 3DEP (30m at the equator, ~10^3 pixels), and
3. The raster pixel dimension changes as a function of latitude, meaning the precision of one (integer) number of pixels increases with increasing latitude.

The smallest number of upstream accumulation cells in the BCUB dataset is 1507 which is a result of the projected grid dimension decreasing with increasing latitude (30m raster pixel would yield a minimum threshold of 1111 pixels). In the coarser resolution, this would represent 170 pixels owing to the factor of 9 area difference between the 30 and 90m grid dimensions. This number of pixels represents the worst case scenario where rounding to an integer number of pixels represents 0.6% rounding error. The BCUB dataset excludes vestigial headwater points, so only pour points (confluences) within 0.6% difference from the area threshold should be affected (included/excluded) between resolutions. Approximately 0.25% of the basins in the BCUB are less than 1.01km^2 representing a 1% deviation from the minimum basin area.

Corresponding manuscript edits:

- The effects of changing resolution are highlighted in sections 2.2.2 and 2.2.3, in particular in figures 7 to 10.
- We did not derive the full attribute set from the lower resolution DEM but the full replication code is provided which can be used for comparisons between datasets and methods.

References:

1. Hrachowitz, Markus, et al. "A decade of Predictions in Ungauged Basins (PUB)—a review." Hydrological sciences journal 58.6 (2013): 1198-1255.
2. Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., and Poulin, A.: “A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds”, Scientific Data, 7, 1-12, 2020.
3. Kratzert, Frederik, et al. "Caravan-A global community dataset for large-sample hydrology." Scientific Data 10.1 (2023): 61.