BCUB - A large sample ungauged basin attribute dataset for British Columbia, Canada.

Daniel Kovacek[1] and Steven Weijs[1]

[1]Department of Civil Engineering, University of British Columbia, 2002 - 6250 Applied Science Lane, Vancouver, V6T 1Z4, BC, Canada

Correspondence: Daniel Kovacek (dkovacek@mail.ubc.ca)

Abstract. The British Columbia Ungauged Basin (BCUB) dataset is an open-source, extensible dataset of attributes describing terrain, soil, land cover, and climate indices of over one million ungauged ~~basins~~ sub-basins in British Columbia, Canada including trans-boundary regions. The ~~basin~~ attributes included in the dataset follow those found in the large sample hydrology literature for their association with hydrological processes. The BCUB database is intended to support water resources research and practice, namely monitoring network analysis studies, or hydrological modelling where basin characterization is used for model calibration. The dataset ~~,~~ and the complete workflow to collect and process input data, to derive stream networks, ~~delineate basins, and to extract basin attributes~~and to delineate sub-basins and extract attributes, is available under a Creative Commons BY 4.0 license. The DOI link for the BCUB dataset is https://doi.org/10.5683/SP3/JNKZVT (Kovacek and Weijs, 2023).

1 Introduction

Spatial datasets available for geoscience research and practice are increasing in size, scale, resolution, and variety. Advances in the capture and processing of remote sensing data have in recent years led to open-access publication of continental and global scale geospatial datasets at high resolution ~~(U.S. Geological Survey, 2022; Gleeson, 2018; Latifovic et al., 2010; Lehner et al., 2021; Thorn~~ (U.S. Geological Survey, 2022; Huscroft et al., 2018; Latifovic et al., 2010; Lehner et al., 2021; Thornton et al., 2021). We are well into the age of high quality, open-access geospatial data anticipated by Hrachowitz et al. (2013) following the decade of prediction in ungauged basins (PUB).

By contrast, the streamflow monitoring network in Canada has contracted over the last three decades. Based on the HYDAT dataset accessed at Environment Canada's national water data archive, the number of streamflow observation locations across Canada peaked in the order of 2300 in the 1980s, and reduced to roughly 1700 in 2022 (on average per day). According to surface water monitoring density standards developed by the World Meteorological Organization (WMO) (via Coulibaly et al. (2013)), nearly 90% of Canada's terrestrial area is under-monitored, and almost 40% is classified as ungauged. In general this trend holds for the province of British Columbia (BC), where outside of a few small regions in the south it is predominantly classified as ungauged or poorly gauged (Coulibaly et al., 2013).

The streamflow data used in a wide range of research and practice today comes from monitoring networks built over many decades, highlighting the significant lag between ~~research aims of the present and monitoring layout decisions of the past~~monitoring objectives of the past and information needs of the present. Monitoring network decisions today must anticipate information needs decades into the future.

Recent deep learning (DL) approaches to regional hydrological modeling use large sample datasets to infer relationships between climate input forcings and streamflow, and model performance ~~improves~~ has been shown to improve when training incorporates static ~~basin~~ catchment attributes (Kratzert et al., 2019). DL models benefit from training datasets (streamflow monitoring networks) representing ~~basins~~ catchments that are diverse in geographic, hydrologic, and geophysical attributes, yet there is no clear consensus on how to evaluate networks in terms of diversity of attributes (Gauch et al., 2021). Increasing monitoring network diversity may be as simple as expanding the monitoring network according to the uniqueness of place described by Beven (2000)~~, or~~ . Alternatively, a different approach ~~can be to define a~~ involves defining the much larger set of ungauged ~~basins~~ catchments and their hydrologically relevant attributes to use as a basis ~~of~~ for comparison.

The vast and growing amount of geospatial information ~~generated~~ available today requires considerable data assimilation effort to support specific research questions. A large, catchment-based dataset of geophysical attributes could support other disciplines that use ~~basin~~ attributes at the catchment level, for example in understanding changing water temperature and its effect on fish habitat (Daigle et al., 2017), or likewise for water quality monitoring in evaluating human-induced concentrations of toxic contaminants in fish (Scholes et al., 2016).

Water resource management decisions are typically made at the catchment level, so research and practice may be well served by datasets that are catchment-based, diverse in characteristics, and large in size and scale to reflect the scale-dependency of physical processes governing the rainfall-runoff response (Arsenault et al., 2020).

1.1 Motivation

The monitoring deficit of a region can be addressed by simply adding more stations, or under resource constraints optimal network arrangements can be approximated based on models trained on existing streamflow monitoring records, combined with information about unmonitored locations (Mishra and Coulibaly, 2010; Werstuck and Coulibaly, 2017, 2018). If large sample datasets improve predictability in ungauged locations by learning from diversity (Addor et al., 2017), a basis is needed to compare the existing monitoring network against the greater region it is intended to represent in relevant hydrological terms. The British Columbia Ungauged Basins (BCUB) (Kovacek and Weijs, 2023) is designed to be a dataset which i) uses only open access data sources that are continuous and complete over the study region, ii) is derived from the highest resolution DEM available to ~~include smaller basins left out of other~~ cover the range of catchment areas represented in large sample hydrology (monitored catchment) datasets, iii) is published under an open-source license, iv) is extensible both spatially and dimensionally to enable integration of new information as it is published, and v) is published with the full replication code based on widely used open-source libraries. Several existing datasets were reviewed for the desired qualities listed above, and for their potential to support research in network optimization, prediction in ungauged ~~basins~~catchments, and water resources more generally.

1.2 Related datasets

The BC Freshwater Atlas (FWA) (Gray, 2010) is the definitive source of freshwater feature mapping for British Columbia (BC). It contains roughly 3 million polygons representing the province-wide set of 1^{st} order fundamental component watershed units with a reference system designed to facilitate aggregation into larger watershed assessment units. The FWA dataset is strictly limited to the administrative bounds of BC, cutting off many important trans-boundary basins at borders. Since the dataset is primarily hydrographic, it does not include static ~~basin~~ catchment attribute information commonly used in rainfall-runoff modelling. The FWA is provided with an open-use license, but the code used to derive the dataset is to our knowledge unpublished, and as such it isn’t readily replicable or extensible with consistent input data and methodology.

The National Hydrographic Network (NHN) (Geobase, 2004) contains a hydrographic feature set similar to the BC FWA. It covers all of Canada and includes trans-boundary basins along the US border, but the geometries are organized in Work Unit Limits (WULs) which ~~do not represent~~ break up complete basins. The watershed attributes are similarly limited, and the code used to derive the geometries is to our knowledge unpublished.

HydroSHEDS is a dataset for global-scale applications featuring river networks, watershed boundaries and other hydrological features derived from the NASA Shuttle Radar Topography Mission (SRTM) DEM for most of North America at a resolution of roughly 90m. At latitudes $> 60°$ North, corresponding to the northern border of BC with the Yukon territory, HydroSHEDS ~~basins~~ catchments are derived from more coarse ($\approx 500m$) Hydro1k (Wickel et al., 2007) elevation data. Attributes derived from distinct elevation data sources are difficult to compare as discussed in subsection 2.2, as the stream networks (and catchment boundaries) are unique to a DEM source and to the data processing methodology (Datta et al., 2022). Studies using the HydroSHEDS dataset typically exclude ~~basins~~ catchments smaller than 100 km^2 (Guth, 2011; Zhang et al., 2020; Kratzert et al., 2023).

Large sample hydrology (LSH) datasets typically specify lower bounds on catchment area to filter out small basins due to uncertainty in basin delineation (Arsenault et al., 2020), ~~and~~ to ensure parameters are derived from sufficiently large samples (Guth, 2011), though ~~explicit justification of~~ quantitative support for a particular threshold is generally not provided. The HYSETS dataset (Arsenault et al., 2020) includes a caveat for attributes describing basins smaller than 50 km^2, representing nearly one third of the dataset. The uncertainty associated with such a large segment of the dataset (and the monitoring network it represents) highlights a gap that can be addressed in part with continuous and complete DEM coverage at greater resolution. ~~The accuracy of stream network delineation improves with increasing DEM resolution (Tarolli and Dalla Fontana, 2009).~~

A large and diverse set of ungauged locations and associated attributes is sought to represent the decision space for network analysis and optimization, and more generally to support water resources research where catchment-based geospatial attributes are relevant.

1.3 British Columbia Ungauged Basin (BCUB) Database

The BCUB database contains a wide array of attributes describing the terrain, land cover, soil permeability and porosity, and climate of over 1.2 million ~~basins.~~ (sub-)basins. We use the term ’basin’ to refer to the local watershed of any confluence

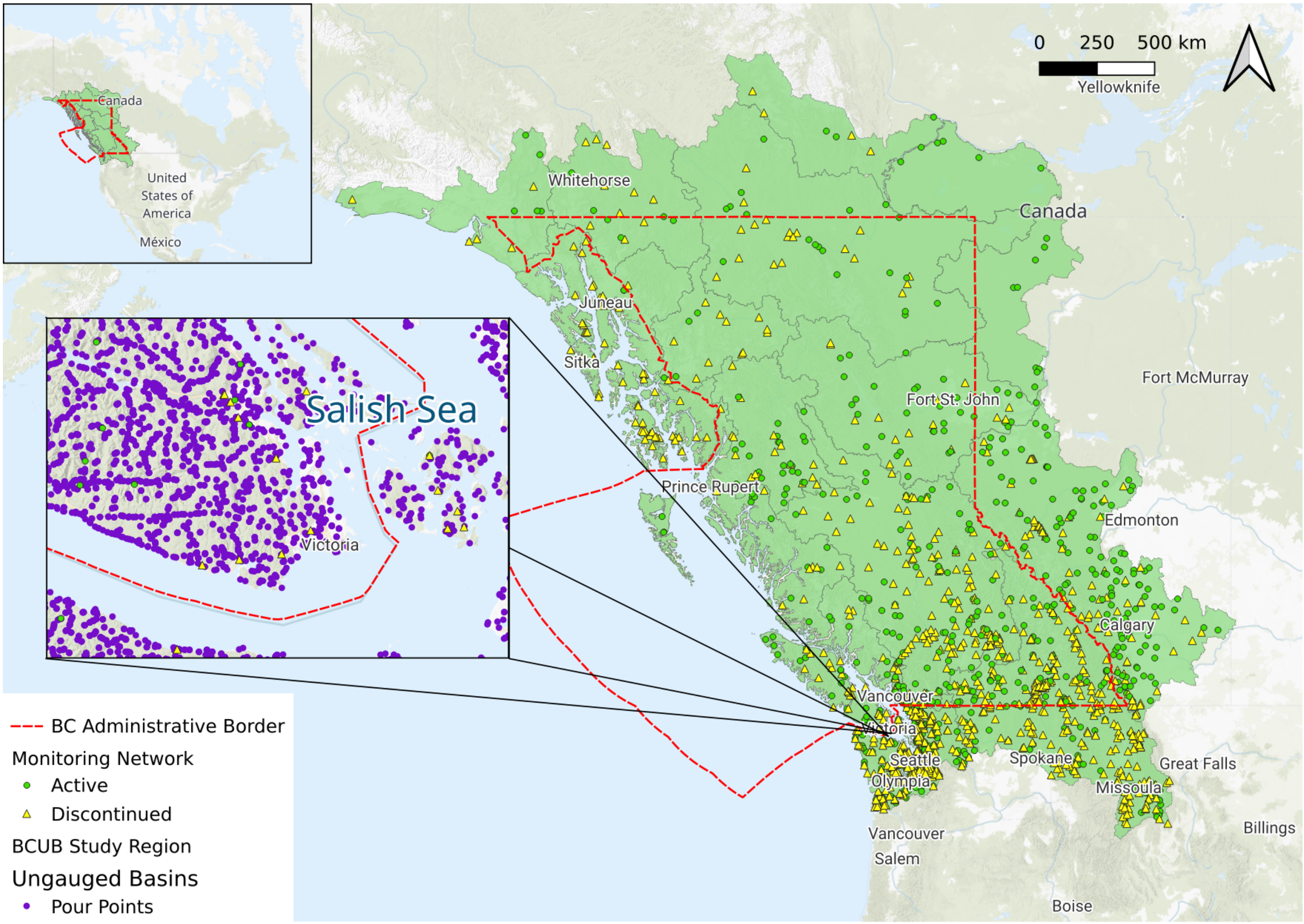

Figure 1. The study region (right) expands beyond the British Columbia administrative border to capture trans-boundary ~~basin~~ regions. Active and discontinued streamflow monitoring stations (those included in ~~Arsenault et al. (2020)~~HYSETS (Arsenault et al., 2020)) are sparse and unevenly distributed as shown in the main figure at right, and the ~~detail~~ inset detail shows ~~a sample of~~ the high density of pour points (purple) defining catchments in the BCUB dataset. (basemap from © MapTiler © OpenStreetMap contributors)

or outlet in a stream network, including individual upstream branches and their combination. Figure 1 shows the ~~ungauged basin~~ pour points representing the BCUB dataset, and the streamflow monitoring stations from the HYSETS dataset (Arsenault et al., 2020) that ~~fall~~ lie within the study region. The study region represents any terrestrial area within or upstream of any point within the BC administrative boundary (red dashed line in Figure 1), plus a buffer to include trans-boundary ~~basins~~ catchments and to mitigate the edge selection bias of optimal sensor placement in random fields (Hershfield, 1965; Rouhani, 1985; Krause et al., 2006).

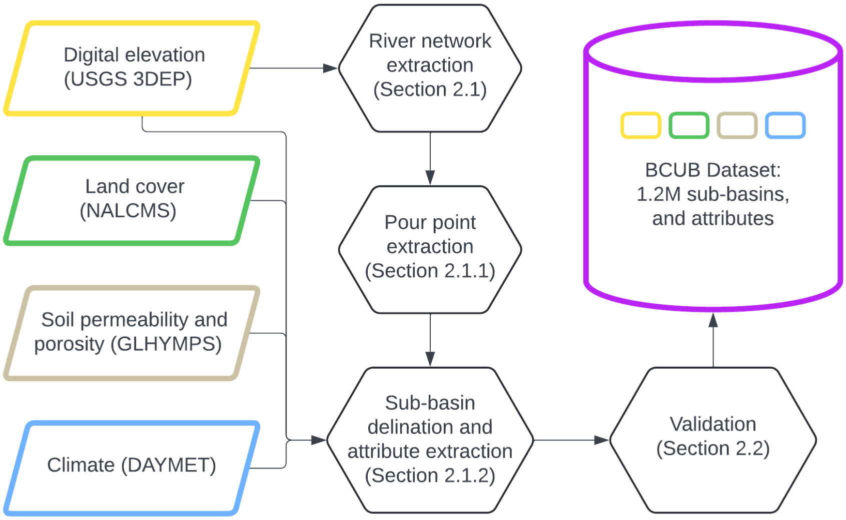

Figure 2. Schematic of the BCUB development pipeline, from retrieving input datasets from external sources to creating a final database of sub-basins and their representative catchment attributes.

The attribute set describing each ~~basin~~ sub-basin follows the HYSETS dataset as much as possible and includes select additional climate indices following the Camels dataset (Addor et al., 2017) to demonstrate how derived parameters can be added to the dataset. Three sets of land cover indices from the North American Land Change Monitoring System (NALCMS) (Latifovic et al., 2010) ~~associated with~~ representing 2010, 2015, and 2020 are included to ~~facilitate evaluation of~~ support questions about land cover change at the basin level as called for by Addor et al. (2020). An example plot showing forest cover change between 2010 and 2020 is shown in section 3.

Following Wilkinson et al. (2016), to support knowledge discovery, innovation, and integration of data and methods in subsequent work, both the data and the code used to generate the data are openly available. The code is provided not to champion a particular method, but to highlight the nuance involved in developing large sample datasets that for brevity and clarity are ~~typically~~ generally left out of dataset description papers. There are no stochastic elements in the methodology, yet there are a large number of methodological choices that yield distinct outcomes. Providing the complete code at minimum aims to be explicit about these choices.

2 Data & Methods

2.1 Data collection and pre-processing overview

~~Static basin attributes in the BCUB dataset were extracted~~ Attributes of ungauged basins were clipped from the digital elevation, land cover, ~~and soilgeospatial layers~~ soil, and climate geospatial datasets described in Table 1 ~~using basin polygons as clipping~~

Table 1. Summary of catchment attribute source data.

Dataset	Attributes	Source
USGS 3DEP[1]	**Terrain**: area, elevation, aspect, slope	(U.S. Geological Survey, 2022)
GLHYMPS[3]	**Soil**: porosity, permeability	~~(Gleeson, 2018)~~ (Huscroft et al., 2018)
NALCMS[2]	**Land cover (2010, 2015, 2020)**: forest, shrubs, grassland, wetland, crops, urban, water, snow and ice	(Latifovic et al., 2010)
DAYMET[4]	**Climate (daily estimates, 1980-2022)**: precipitation, temperature, snow water equivalent, vapour pressure, shortwave radiation	(Thornton et al., 2022)

1. 3DEP: 3D Elevation Program, U.S. Geological Survey,

2. NALCMS: North American Land Change Monitoring System, accessed at http://www.cec.org/north-american-land-change-monitoring-system/

3. Global Hydrogeology Maps.

4. Gridded daily climate estimates on a 1-km Grid for North America, Version 4. https://daymet.ornl.gov/

~~masks. Basin~~ through a data preparation and processing pipeline described in Figure 2. Individual catchment polygons were delineated from the set of pour points in the stream network representing river confluences. The stream network was derived from the 1 arc-second (30m at the equator) resolution USGS 3DEP (U.S. Geological Survey, 2022) digital elevation model (DEM) using the open-source software library Whitebox (version 2.3) (Lindsay, 2016). Streams are defined by a minimum upstream accumulation of 1 km^2 to match the smallest monitored catchment in the HYSETS dataset.

The study region was divided into complete basin sub-regions (no surface inflow across boundaries) as shown in Figure 3 (right) assembled from HydroBASINS (Lehner et al., 2021) data to simplify the automated ~~basin~~ sub-basin delineation and attribute extraction work flow. ~~Additional details about the data collection and pre-processing steps for generating the BCUB basin polygons are provided below.~~ The data processing pipeline is described as follows:

1. **Define study region and sub-regions**: Level 5 and 6 watersheds from the HydroBASINS dataset were used as a first approximation to break the study region into smaller components for memory management in data pre-processing. Study region bounds were refined by deriving the covering set of basins in each region independently, see subsubsection 2.2.1 for more detail about the treatment of region bounds.

2. **Retrieve DEM data**: The study region bounding box was used to download the covering set of digital elevation tiles from the USGS 3D Elevation Program (U.S. Geological Survey, 2022). In addition, lower resolution (90m) DEM tiles from EarthEnv DEM90 (Robinson et al., 2014) were used in the data validation analysis presented in subsection 2.2.

3. **Pre-process DEM raster**: Hydraulic conditioning of the DEM, including depression filling, resolving flats, computing flow direction and accumulation, and stream network extraction were processed using the open-source geospatial analysis software Whitebox (Lindsay, 2016).

4. **Define and filter pour points**: Pour points define the outlet of each catchment and their precise location is specific to the input DEM and pre-processing steps. Each ungauged catchment is delineated from a pour point defined by the

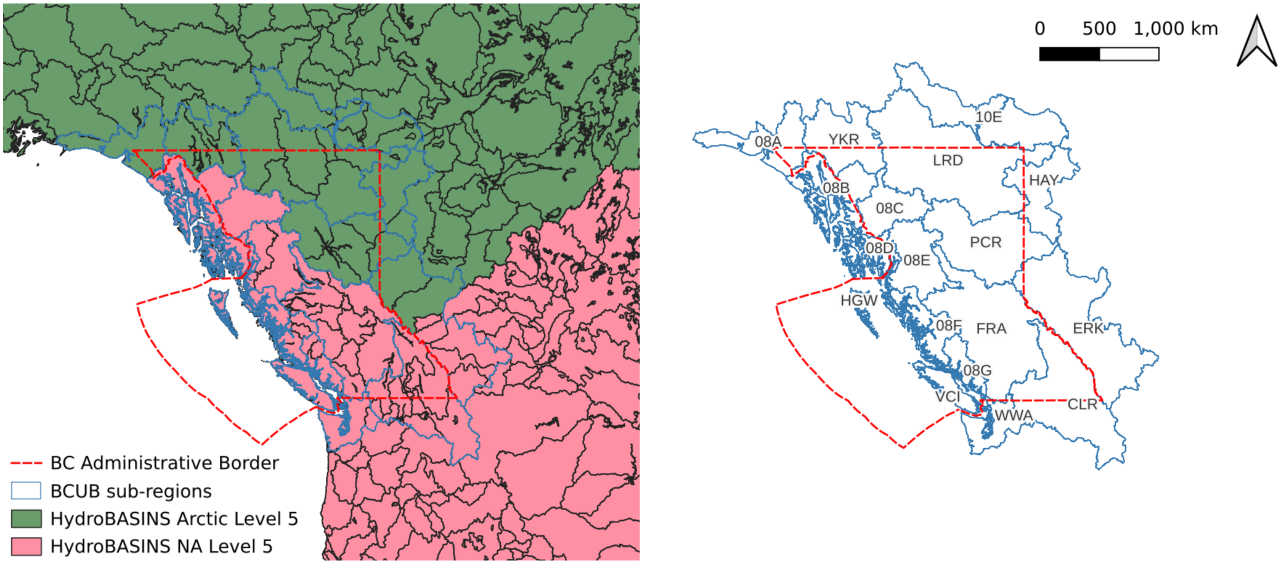

Figure 3. At right, the study region is divided into complete watershed sub-regions (encoded in the "region_code" parameter) by merging level 5 & 6 HydroBASINS polygons to cover the BC boundary~~and~~. The study region extends beyond the administrative border of BC to include trans-boundary basins and ~~include~~ a minimum buffer of $\approx$ 100 km. The purpose of merging complete watershed regions is to manage computational resources the DEM pre-processing → ~~basin~~ sub-basin delineation → attribute extraction pipeline.

stream network. Lake polygons from HydroBASINS were used to filter out pour points within lakes. Points are flagged (*in_perennial_ice*) where the 2020 NALCMS land cover classification is perennial ice and snow.

5. **Catchment delineation**: Catchment polygons were derived from sets of input pour point coordinates using the "Unnest-Basins" function in Whitebox.
6. **Attribute extraction**: Catchment polygons were used as clipping masks to capture representative values from the various geospatial layers. Attribute indices were aggregated from raster and vector layers as described in Table 2.

Additional detail about pour point selection, catchment attribute extraction, and data processing follows.

2.1.1 Pour point set selection

The ~~basins~~ sub-basins in the BCUB database are delineated from a subset of raster cells representing the stream network. The set of pour points points used for ~~basin~~ catchment delineation is called the *candidate monitoring location* (CML) set. By limiting the CML set to river confluences, the number of ~~basins to process reduces to~~ polygons to process is reduced to $<$ 5% of the complete set of stream network cells. Since changes in upstream accumulated area are small along reaches between confluences, and by extension changes in the hydrologic properties of the ~~basin, are smallalong reaches between~~

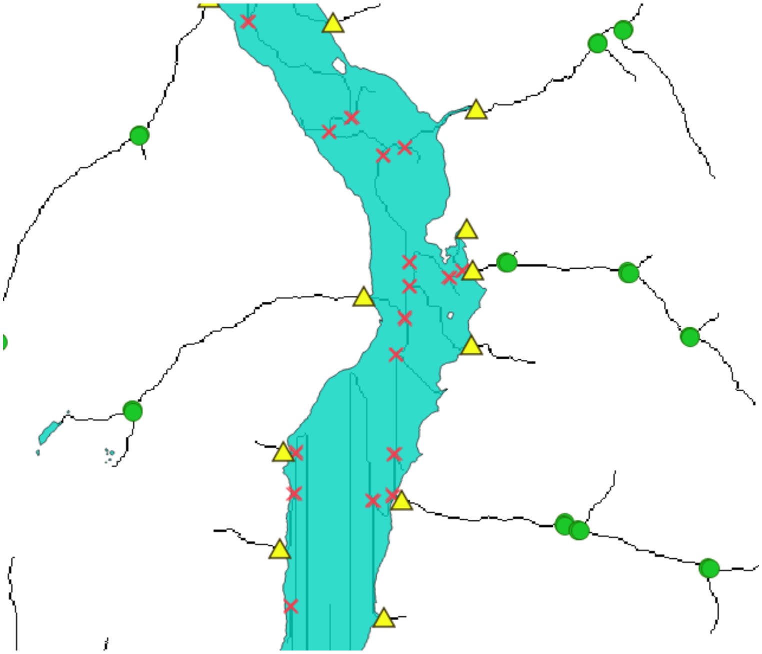

Figure 4. Example of river confluence (green circles) where spurious confluence points within lakes are excluded (red "x") and river-lake confluences are added (yellow triangle).

~~confluences~~sub-basin are small, eliminating these points reduces redundancy and data processing~~, and reduces the decision space for subsequent network optimization analysis~~.

The CML set is defined by the following ~~selection~~ criteria:

1. **Confluences**: stream cells with more than two neighbouring stream cells (8-direction grid), where the flow direction of more than one neighbouring stream cell is pointed toward the target cell, and
2. **River outlets**: intersections of river network lines with ~~with~~ ocean coastline, major ~~basin~~ regional watershed outlets at the study region boundary, and ~~intersections~~ confluences with lakes where the upstream contributing area is at least 1 km^2.

Stream confluences within lakes were excluded from the pour point set, as illustrated in Figure 4 where a red "x" denotes a spurious confluence within a lake, a yellow triangle represents the location where a river drains into a lake. Green circles represent ~~upstream branches and their combination individually~~confluences and individual upstream branches.

The ~~stream network is defined by raster cells with a minimum upstream accumulation of 1 km^2. The~~ headwaters mapped in the stream network ~~raster~~ are simply a vestige of the minimum area threshold (1 km^2) used to define a stream network, so they are ~~not included in~~ excluded from the pour point set. Accurate headwater identification (network extent mapping) requires a more rigorous approach to address uncertainty related to stream permanence (Shavers and Stanislawski, 2020). Mutzner et al. (2016) found classical (i.e. cumulative drainage area) threshold approaches do not capture spatial variability of headwater drainage networks in mountainous regions compared to detailed field survey mapping, and statistical methods are likewise unable to resolve local topography to accurately map headwater streams at low-resolution. Further discussion of uncertainty ~~in stream networks~~ is provided in subsection 2.2.

2.1.2 ~~Attribute extraction~~Sub-basin delineation and notes on attributes

A ~~basin polygon was delineated~~ catchment boundary polygon was generated for each pour point in the CML set using the "unnest basins" function in the Whitebox software library (Lindsay, 2016). ~~Basin attributes~~ Attributes were derived for each ~~basin~~ sub-basin by i) using ~~basin~~ the polygons as raster clipping masks, and ii) spatial intersection of the ~~basin~~ polygon and geospatial raster and vector data in PostGIS (PostGIS Project Steering Committee and others, 2018). ~~Basin attribute descriptions are provided in , and metadata attributes are described separately in . Attribute values are~~

Attribute values were computed using the geometric mean of the raster pixel values contained in basin polygons in the case of soil permeability, the circular mean in the case of slope aspect, the fraction of total area in the case of land use, and the spatial mean for all other attributes. ~~Additional pre-processing steps for the Daymet climate data~~ Physical attributes are described in Table 2, and metadata attributes are described in Table 3.

~~Two~~ Several binary attributes are included in the attribute set ~~. A soil flag~~ to represent uncertainty in geometry and value estimates. A ‘soil_flag‘ value of 1 indicates that the ~~sum of soil polygon areas clipped with the basin mask~~ clipped soil data differs from the catchment polygon area by more than ~~10% , to reflect where gaps exist~~ 5% to indicate gaps in the GLHYMPS ~~vector set. A permafrost flag~~ (soil) data. A ‘permafrost_flag‘ value of 1 represents the presence of permafrost in the basin.

~~Expansion of the study region or addition of new attributes can be accomplished by following the processing methodology in the code repository provided. Four parameters derived from the Daymet daily precipitation data are processed inthe code provided do demonstrate how computed parameters can be added from existing input data. The examples follow the Camels dataset Addor et al. (2017) and include:~~

~~**Low precipitation frequency**: frequency of days where precipitation < 1 mm day^{-1}), **Low precipitation duration**: average duration of low precipitation events, or the number of consecutive low precipitation days < 1 mm day^{-1}, **High precipitation frequency**: frequency of days where precipitation is ≥ 5 times the mean daily precipitation, and **High precipitation duration**: average duration of consecutive high duration events, number of consecutive high precipitation days $\geq$~~ A value of 1 for the ‘in_perennial_ice‘ flag represents a pour point location where the land cover classification is "perennial snow and ice" as defined by (Latifovic et al., 2010). A ‘geometry_flag‘ value of 1 represents a catchment intersecting or touching an uncertain area along the region boundary whose area is >= 5~~times mean daily precipitation.~~ % of the catchment area, as described in subsubsection 2.2.1.

2.1.3 Data processing notes

Beyond data sources, the offline approach of deriving ~~basins~~ sub-basins from source data and writing code to process attributes was adopted despite the elegant online polygon aggregation and processing approach demonstrated by Kratzert et al. (2023) in developing the Caravan dataset with use of Google Earth Engine (GEE) (Gorelick et al., 2017). Such an approach is preferable from the perspective of standardized methods of ~~basin~~ catchment attribute extraction, but for our target of ungauged ~~basins~~ catchments it does not eliminate the need for DEM pre-processing to generate stream networks, for filtering and extracting pour points, or for ~~basin~~ sub-basin delineation. These steps represent a substantial portion of the ~~basin~~ attribute extraction workflow, and ~~the~~ what remains to process with GEE is still subject to usage limits, namely for processing the very large set of polygons, even considering an aggregated polygon approach.

A benefit of the offline approach is generating ~~set of basin polygons from higher resolution DEM~~ a set of sub-basin polygons from the highest resolution DEM available that is continuous and complete, and ensuring that basin polygons match the DEM source from which terrain attributes are derived.

Expansion of the study region or addition of new attributes can be accomplished by following the processing methodology in the code repository provided. Four parameters derived from the Daymet daily precipitation data are processed in the code provided do demonstrate how computed parameters can be added to the BCUB from existing input data. The examples follow the Camels dataset Addor et al. (2017) and include:

1. **Low precipitation frequency**: frequency of days where precipitation < 1 mm day^{-1}),
2. **Low precipitation duration**: average duration of low precipitation events, or the number of consecutive low precipitation days < 1 mm day^{-1},
3. **High precipitation frequency**: frequency of days where precipitation is ≥ 5 times the mean daily precipitation, and
4. **High precipitation duration**: average duration of consecutive high duration events, number of consecutive high precipitation days ≥ 5 times mean daily precipitation.

2.2 Technical Validation

The large number of ~~basins~~ geometries in the BCUB dataset requires an automated approach to ~~stream network validation and the basin~~ validate the sub-basin polygons used to capture ~~basin~~ attributes. The representativeness of ~~basin~~ attributes is a function of the accuracy of the stream network derived from DEM. Higher resolution DEM can better resolve lower-relief topographic features resulting in better basin delineation performance, particularly for small basins (Zhang and Montgomery, 1994; Tarolli and Dalla Fontana, 2009; Woodrow et al., 2016).

It is important to emphasize that the 1 km^2 minimum ~~basin~~ drainage area threshold introduces significant uncertainty in the accuracy of the smallest ~~basins, and in basins~~ sub-basins, and those where topographic relief is low. Detailed validation of stream network accuracy is left to future work that the BCUB is intended to support, and validation of the smallest ~~basins~~ sub-basins used in studies is left to the user. Next we discuss indirect attribute validation methods, and limitations of the dataset and methods. ~~The code to replicate the figures in this section is provided in the associated Github repositoryin the "validation" folder~~

2.2.1 Region boundary treatment

While the region polygons assembled from HydroBASINS are a helpful tool for organizing the data processing pipeline, the resulting bounds are different from those produced by independently delineating basins from the 1 arc-second DEM used in this study. These differences are comparable in size to the smallest sub-basins in the BCUB dataset, introducing uncertainty into the attributes of any catchment whose boundary touches or intersects them. Boundary deviations are defined as i) gaps between region bounds where the DEM does not resolve an outlet, and ii) boundary overlaps between regions with shared boundaries.

The Caravan dataset (Kratzert et al., 2023) clearly describes the issue with aggregating attributes from catchment boundary polygons that do not precisely align with the HydroBASINS polygons. By independently deriving the region bounds from a single continuous DEM source (1 arc-second USGS 3DEP), we avoid the problem of misalignment with HydroBASINS polygon. This process does not guarantee perfect alignment of region bounds, but the mean size of deviations is significantly reduced.

The edge detail inset Figure 5 shows an example segment of region boundaries aggregated from HydroBASINS (blue dashed line) compared to those derived from the USGS 3DEP (1 arc-second) DEM. In Figure 5, the purple (Peace, PCR) and green (Fraser, FRA) areas represent the (BCUB region) boundaries delineated from the 1 arc-second DEM. White areas are gaps that remain following the iterative boundary definition process described below.

To avoid restricting the catchment boundary delineation by the clipping mask, a (5 km) buffer was applied to the region boundaries aggregated from level 5 and 6 HydroBASINS polygons. The buffered polygons were used as clipping masks on the DEM before deriving the covering set of polygons (catchments) for each region. The covering set is defined as the smallest number of non-overlapping polygons covering a region. The exterior edges (of the union of intersecting geometries) were checked to verify that they do not touch the edge of each buffered region polygon. Where the edges intersect, the buffer (DEM clipping mask) was manually expanded in QGIS and the process repeated until the buffer was sufficient, i.e. the covering set of basins does not touch the edge of the clipping mask. The use of a buffer produces small peripheral catchments draining to adjacent region basins, and these are excluded by identifying that they are completely contained by the clipping mask of the adjacent regions.

Delineating region boundaries independently from the HydroBASINS polygons does not yield perfectly shared boundaries, but the resulting deviations are substantially smaller. The distribution of the size of deviations from shared sub-region boundaries is shown in Figure 6. The red series represents differences between the BCUB region bounds and HydroBASINS-derived bounds (median area of $0.13\mathrm{km}^2$), while the blue series represents disagreement (overlaps and gaps) between the BCUB sub-region boundaries (median area $0.025\mathrm{km}^2$). Polygons smaller than 0.01 km2, or 1% of the smallest sub-basin in the BCUB dataset were neglected. The boundary deviation polygons (gaps and overlaps) are included in the code repository.

The uncertainty introduced by missing or overlapping areas along sub-region bounds is addressed in the BCUB dataset in two ways. The ‘geometry_flag‘ attribute indicates that a catchment polygon intersects or touches an uncertain region bound if the total uncertain area represents at least 5% of the catchment area. Where catchments derived from distinct basin outlets overlap, either catchment may overestimate the area, and where an area is not covered by any basin but is not necessarily endorheic, either bordering sub-basin may underestimate the catchment area. Where a catchment polygon touches or intersects with an

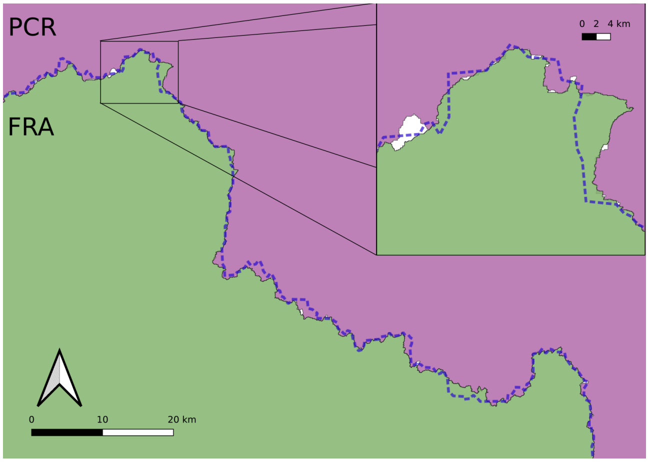

Figure 5. An example edge detail of the shared boundary between the Peace (PCR, purple) and Fraser (FRA, green) basins. The blue dashed line represents the HydroBASINS bounds while the coloured areas represent sub-regions delineated independently from USGS 3DEP DEM.

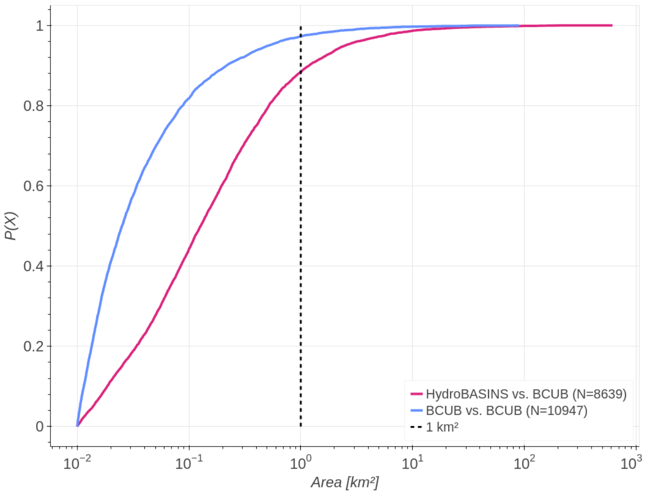

Figure 6. Distributions of geometric deviations (uncertain edges) between shared sub-region boundaries from HydroBASINS-based region polygons (red series, median area $0.13\,\mathrm{km}^2$) and the improvement from deriving region bounds (blue, median area $0.025\,\mathrm{km}^2$).

uncertain boundary, the size of the uncertain area is represented by a positive integer value to indicate potential overestimation (‘inside_pct_area_flag‘) or underestimation (‘outside_pct_area_flag‘) of the catchment as a percentage of the catchment area. The purpose of including these quantities is to identify and express the significance of uncertain catchment bounds.

2.2.2 Vestigial effects of DEM resolution

In addition to the ~~process of hydraulic conditioning~~ hydraulic conditioning process for stream network derivation, the grid representation of elevation introduces vestigial artifacts in the representation of basins~~and by extension in the extraction of basin attributes~~, and consequently, catchment attribute estimates.

The stream network derived from DEM does not capture permanent water bodies, resulting in spurious river confluences. These vestigial confluences were excluded by using the lakes geometry layer from HydroBASINS as a mask, as described in subsubsection 2.1.1. Since HydroBASINS is derived from different sources, ~~geometries~~ hydrographic features do no align exactly with the stream network we derived from the 1 arc-second DEM.

The disk space required to store a polygon is a linear function of the number of vertices defining it, and the precision of geographic coordinates describing the geometry. The ~~basin~~ sub-basin polygons are simplified (using the Shapely library (Gillies, 2021) "simplify" function) using a tolerance equal to one-half the diagonal length of the raster pixel resolution. Simplifying (or smoothing) ~~basin polygons is~~ polygons represents a trade-off between reducing the disk and bandwidth required to store and transmit large sets of ~~basin~~ geometries, and the representativeness of attributes that are captured by intersecting

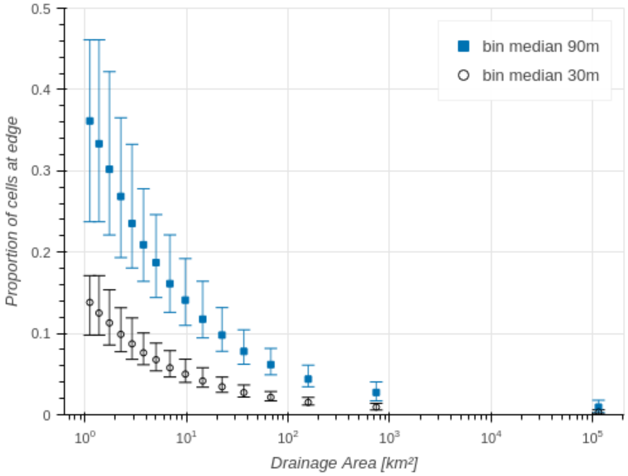

Figure 7. As the ~~basin~~ drainage area decreases, the number edge pixels becomes a significant proportion of the total number of pixels representing the ~~basin~~sub-basin. Points in the above figure represent bin median values based on equiprobable binning ($N \approx 600$ samples per bin), and the whiskers represent the 5 and 95 percentile values for each bin.

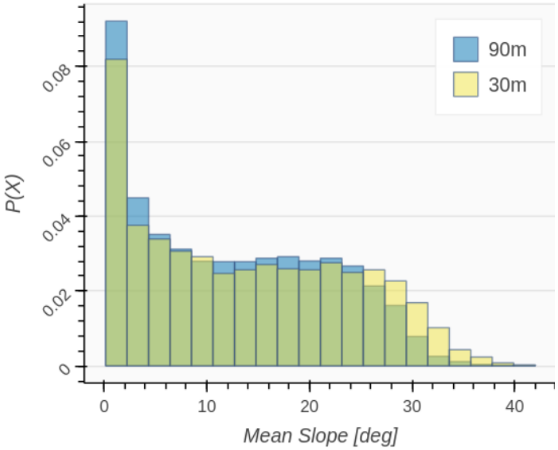

Figure 8. Higher resolution DEM captures greater topographic relief as shown by comparing the distribution of mean slope between 30m (USGS 3DEP) and 90m (EarthEnv) DEM on a random sample of 10,000 basins.

~~basin polygons~~ each polygon with the various geospatial raster layers. The effect of polygon simplification is discussed in more detail in subsubsection 2.2.3.

The set of raster pixels representing each ~~basin are~~ sub-basin is captured using the "crop-to-cutline" function from the open source GDAL library (GDAL/OGR contributors, 2023) which by default captures pixels whose centroid lies within the polygon (pixels are not points, but quadrilaterals). Alternatively the larger set of *intersecting* pixels can be selected by setting the "CUTLINE_ALL_TOUCHED=TRUE" keyword argument. As ~~basin~~ drainage area decreases (or raster resolution decreases), the difference in edge pixel selection method represents an increasing proportion of total pixels which may then yield significant differences in attribute values depending upon the clipping method used. Figure 7 shows that the proportion of edge pixels representing the ~~basin~~ catchment increases with decreasing area~~on a large sample of basin polygons and compare~~, and uses the USGS 3DEP (30m grid at the equator) and EarthENV DEM90 (EENV) DEM (90m grid at the equator) ~~DEM (EENV) on the same sample of polygons. The BCUB is derived from the USGS 3DEP (30m at the equator) DEM, and this exercise highlights~~ to show how the proportion of edge pixels changes with DEM resolution. The purpose of this exercise is to highlight one source of uncertainty introduced by the data processing methodology and ~~suggests at what scale of basin the choice of clipping method becomes significant~~to demonstrate the effect of the clipping method as a function of catchment scale.

Mean ~~basin~~ slope is a widely used attribute ~~(Addor et al., 2017; Alvarez-Garreton et al., 2018)~~in large sample hydrology (Addor et al., 2017; Alvarez-Garreton et al., 2018) to describe the degree of topographic relief of a catchment, defined in Arsenault et al. (2020) as "the average slope when considering the individual elevation differences between tiles" (raster pixels)~~and describes the basin's topographic relief~~. We used ~~the WhiteboxTools Slope gradient function which computes slope for~~ WhiteboxTools to compute the slope of each DEM pixel using a 3rd-order Taylor polynomial fit (Florinsky, 2016) with a kernel size of 5x5 pixels. Mean ~~basin~~ catchment slope increases with increasing resolution because topographic relief is better captured at higher resolution (Zhang and Montgomery, 1994). Figure 8 compares mean ~~basin slope for two DEM sources with~~ slope between 30m and 90m resolution DEM sources, where the higher resolution DEM is able to resolve greater topographic detail. The comparison is based on a random sample of ~~10K basins~~ roughly ten thousand polygons in the BCUB dataset ranging in size from 1 km^2 to 2×10^5 km^2.

The sample of ~~basins~~ sub-basins in Figure 8 shows a bias toward lower calculated mean slope from the lower resolution DEM ~~sources~~ source using the same ~~basin polygon~~ polygon mask to capture pixels. Further interpretation of these differences is left to future work.

2.2.3 ~~Basin~~ Catchment Attributes and Self-Similarity

Mandelbrot (1967) described the measurement of coastline length as a function of the scale of observation, and the lines describing features like catchment boundaries and stream networks also exhibit self-similarity. ~~Basin perimeter~~Perimeter, stream gradient, and shape factors like elongation or compactness are length-based attributes used in many LSH datasets (Arsenault et al., 2020; Klingler et al., 2021; Kratzert et al., 2023). The ~~basin~~ compactness coefficient is defined as the ratio of ~~basin~~ polygon perimeter to the circumference of a circle with equal ~~basin~~ area ((Gravelius, 1914) as cited in (Sassolas-Serrayet et al., 2018)) . Length-based attributes are not comparable without consistent input DEM resolution and data pre-processing.

The difference in catchment boundary lines shown in Figure 9 illustrates why perimeter measurement can vary considerably due to input DEM resolution or ~~basin~~ catchment delineation methodology. ~~Basin perimeter~~ Perimeter is not included in the BCUB attribute set because unless otherwise treated, ~~basin~~ polygons derived from higher resolution DEM will measure a longer perimeter.

~~A large number of polygons used in LSH datasets (HYSETS and Caravan) were revised by~~ In July 2022 the Water Survey of Canada ~~in July 2022 and~~ published updated catchment boundaries representing the majority of the streamflow monitoring network. These updated geometries can be accessed at the (WSC) National Water Data Archive. We found all polygons common to both the HYSETS dataset and this updated polygon set, and computed pairwise comparisons of perimeter lengths. There were 1035 sub-basin polygon revisions that did not meet the similarity criteria, reflecting the difficulty in retrospectively defining streamflow monitoring station locations from historical records (Arsenault et al., 2020).

The sample used for the perimeter comparison includes 715 ~~basins~~ sub-basins where the original and updated polygons were a close match ~~. Similarity was evaluated based on~~ to control for significant changes in the polygon shape. A "close match" is defined as the ratio of intersecting area to union area (Jaccard similarity index) $\geq$ 95%~~to control for significant changes in the polygon shape~~. Figure 10 shows the newer revision polygon perimeter measurements are substantially greater, and ~~that the deviation is not sensitive to the basin~~ the deviation exists independent of spatial scale. This difference highlights the need

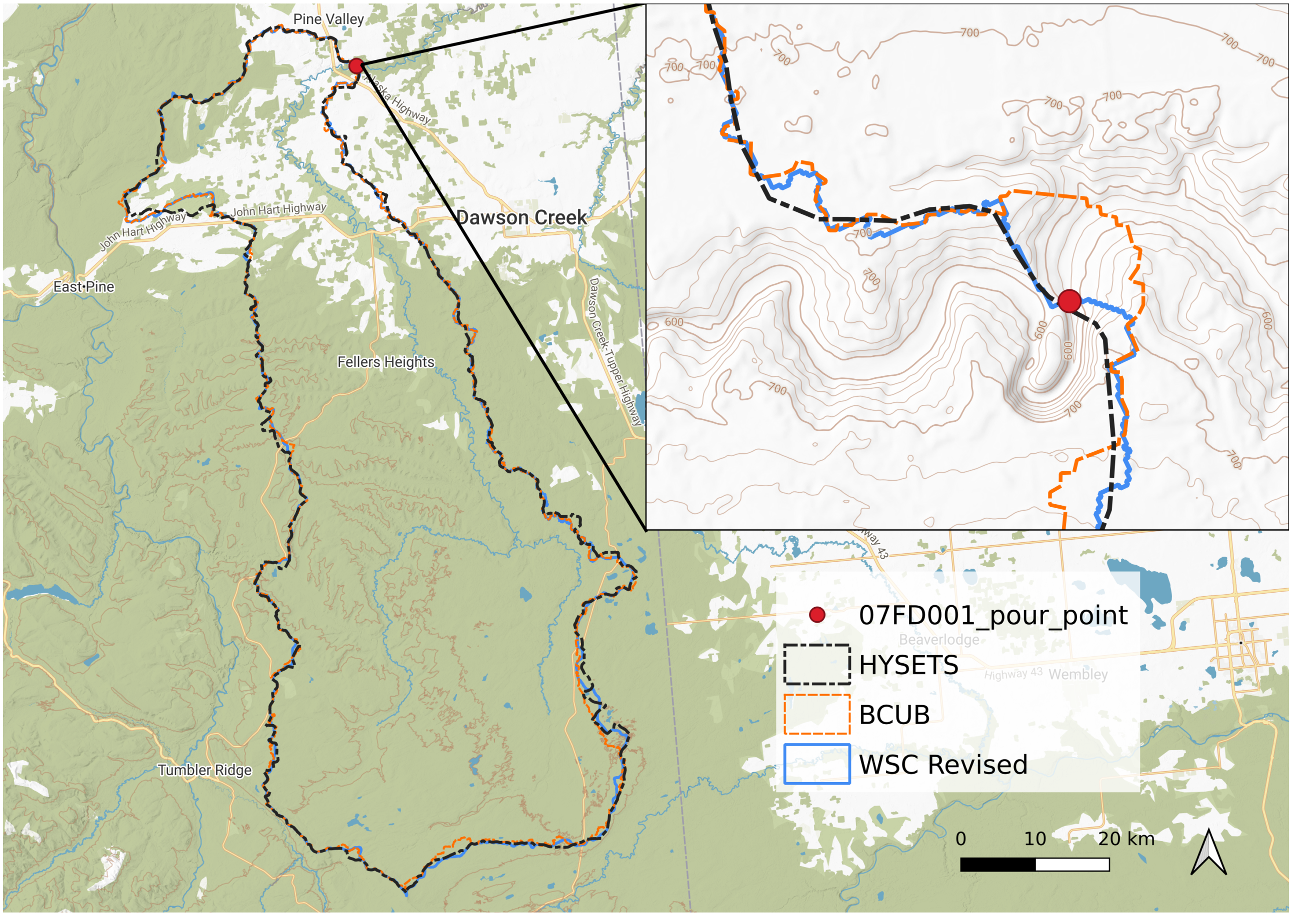

Figure 9. An example edge detail of the same catchment boundary from three different sources where the intersecting area is over 98% of the published value. The HYSETS dataset polygon (back dash-dot line) comes from an earlier revision published by the WSC representing the Kiskatinaw River near Farmington (WSC ID 07FD001), while a recent revision (July 2022) by the WSC (solid blue) shows a distinct difference in polygon edges. The polygon from the BCUB (dashed orange) derived from USGS 3DEP DEM is different from both. (basemap from © MapTiler © OpenStreetMap contributors)

to ensure consistent ~~, continuous~~ input DEM and data processing methodology if length-based attributes are included ~~basin~~ attribute datasets.

~~There were 1035 basin polygon revisions that did not meet the similarity criteria, reflecting the difficulty in retrospectively determining streamflow monitoring station locations from historical records (Arsenault et al., 2020).~~

Average stream gradient is a length-based ~~basin~~ attribute that is a function of both raster resolution and the assumed location of channel head, usually by minimum area threshold. Robinson et al. (2014) calculated mean stream gradient as the ratio of the maximum total elevation change in the basin stream network to the length of the corresponding river reach. Stream length is a function of DEM resolution, and the length of reach is measured from the ~~basin~~ catchment outlet to an uncertain headwater

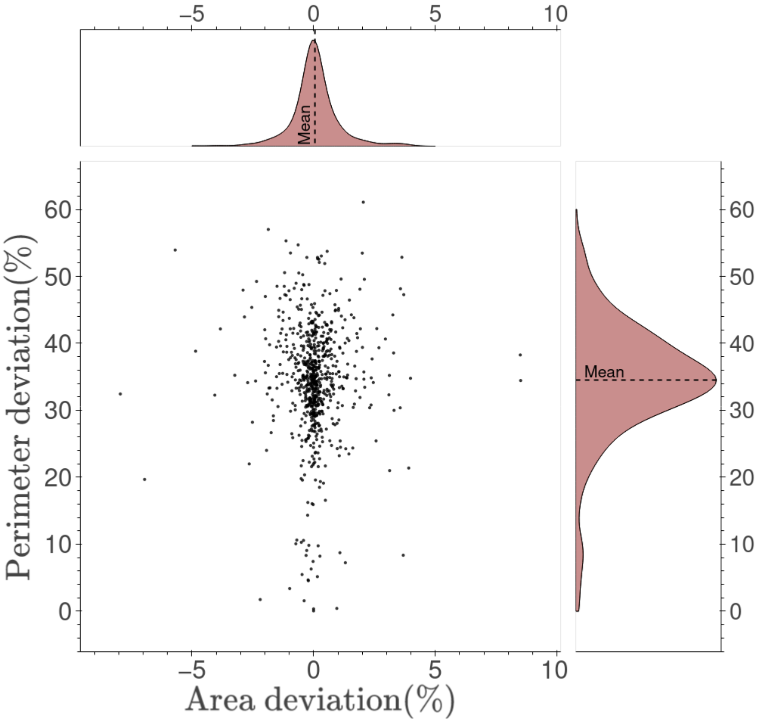

Figure 10. The DEM resolution and the processing methods used to derive ~~basin polygons~~ catchment boundaries affects the measurement of perimeter. Polygons derived from different sources or using different methodologies will yield different values. Comparing sequential revisions of the same streamflow monitoring stations, the perimeter length is significantly different despite the area being ~~roughly~~ nearly constant, and despite a close match between polygons according to a Jaccard Similarity Index match of $\geq 95\%$.

location (Hafen et al., 2020, 2022). In the derivation of the stream network for the BCUB dataset, headwater locations are simply a vestige of the assumed minimum drainage area threshold, and as a result an attribute representing average stream gradient is not included in the BCUB database.

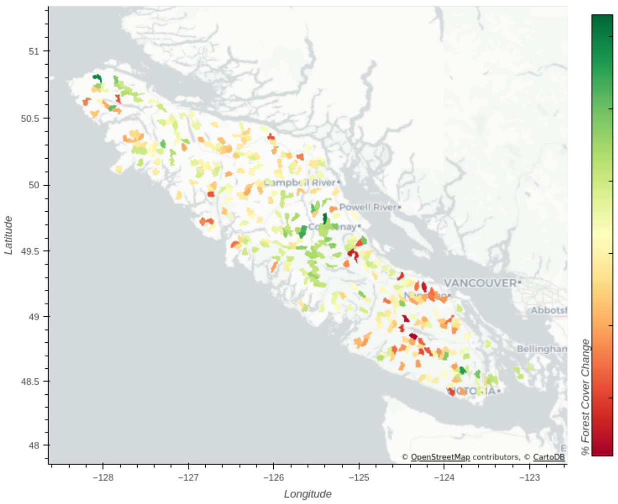

Figure 11. An example visualization using the BCUB dataset maps the percent change in forest cover (as a percentage of the ~~basin~~ catchment area) for ~~basins~~ sub-basins with drainage area between 20 and 25 km^2 on Vancouver Island (VCI). Basemap from © OpenStreetMap contributors.

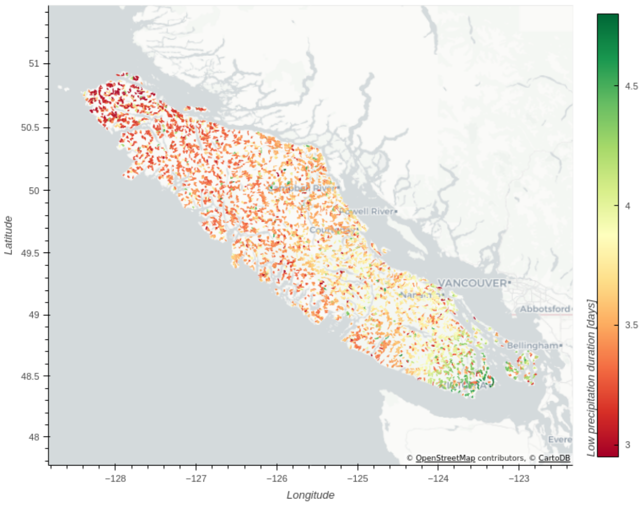

Figure 12. An example visualization using the BCUB dataset maps the mean annual duration of low precipitation (< 0.1mm/day) for ~~basins~~ catchments with drainage area between 2 and 5 km^2 on Vancouver Island (VCI). Basemap from © OpenStreetMap contributors.

3 Usage Notes

It is the hope that the BCUB dataset will serve a wide range of water resource research and practice where catchment-based attributes are integral to the methodology, or perhaps more importantly to express the limits of appropriate use and interpretation. Figure 11 and Figure 12 provide two basic examples of the kind of ~~basin-level~~ sub-basin level querying the BCUB is designed to support. Figure 11 shows ~~basin-level~~ catchment-level changes in forest cover between 2010 and 2020 for basins in the range of 20 to 25 km^2, and Figure 12 shows the mean duration of dry periods (days with less than 0.1 mm rainfall) for ~~basins~~ catchments between 2 and 5 km^2.

Stream networks are unique to the input DEM, and they are affected by the choice of pre-processing steps. The greatest degree of uncertainty is associated with the smallest ~~basins~~ catchments with the lowest topographic relief. Zhang and Montgomery (1994) provides guidance about interpreting features at scales relative to DEM resolution. The representativeness of stream networks, and by extension ~~basin polygons and~~ the attributes captured by polygon ~~masking~~masks generated from stream networks, is an important component of uncertainty analysis and data reliability assessment. This aspect of the analysis is left to future work that the BCUB dataset is designed to support, in particular the lower limit of basin scale that can be supported by 1 arc-second DEM.

4 Code and data availability

The BCUB dataset (Kovacek and Weijs, 2023) is accessible under a Creative Commons BY 4.0 license through the Borealis data repository at https://doi.org/10.5683/SP3/JNKZVT. A summary of the dataset contents and supporting information is presented in Table 4. The sub-basin polygon geometries are provided in the open-source, cross-language Apache Parquet format (https://parquet.apache.org/), which has the convenience of supporting multiple geometries. The Parquet file format is supported by several widely used Python libraries, including Dask (https://docs.dask.org/) and GeoPandas (https://geopandas.org/), and the Arrow package features an interface for the R programming language (https://arrow.apache.org/docs/r/). The dask-geopandas library in Python (https://dask-geopandas.readthedocs.io/) is recommended for performance with large datasets.

The catchment attributes are provided in two forms in the Borealis data repository. The larger form includes catchment boundary, centroid, and pour point geometries. These are saved in the Parquet file format under the ‘basin_polygons‘ folder (select the "tree" view for easier navigation). The Parquet file naming convention follows the sub-region codes shown in Figure 3. A "light" format without geometries is provided in comma delimited format in BCUB_attributes_20240630.csv. Sub-region geometries with their associated codes are provided for reference in BCUB_regions_4326.geojson. Metadata describing the dataset is provided in MetaData.pdf, and additional sub-basin attribute information, including descriptions and sources is provided in the Readme.pdf.

The scripts used to derive the dataset, and the validation results and figures shown in this paper are provided in an open-source Github repository (https://github.com/dankovacek/bcub). The code to replicate the figures in subsection 2.2 is provided in the "validation" folder of the repository. Figures 1 to 3 and 6 were prepared with the QGIS software (QGIS Development Team, 2023), and all remaining figures were created using the Bokeh data visualization library (Bokeh Development Team, 2023) in Python.

In addition, an example guide is provided (https://dankovacek.github.io/bcub_demo/) through a set of Jupyter (Kluyver et al., 2016) notebooks to demonstrate the complete process of data retrieval, pre-processing, sub-basin delineation, attribute extraction, and data product usage. The code to produce Figure 11 using the Parquet file format is demonstrated in the final chapter of the Jupyter book demo, titled "Data Import and Usage Examples".

Author contributions. Daniel Kovacek wrote the code to create the dataset and the Jupyter Notebook tutorials, and wrote the manuscript. Steven Weijs provided research supervision and manuscript review.

Competing interests. The authors declare no competing interests.

Acknowledgements. This study received financial support from the British Columbia Ministry of Environment and Climate Change Strategy (Agreement #TP23EPEMA0031MY). The authors wish to thank the two anonymous reviewers whose feedback improved the quality of this manuscript. The authors also wish to express gratitude to all those contributing to open-source scientific software.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, 2017.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrological Sciences Journal, 65, 712–725, 2020.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., et al.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies–Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5846, 2018.

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.

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrology and earth system sciences, 4, 203–213, 2000.

Bokeh Development Team: Bokeh: Python library for interactive visualization, https://bokeh.org, 2023.

Coulibaly, P., Samuel, J., Pietroniro, A., and Harvey, D.: Evaluation of Canadian National Hydrometric Network density based on WMO 2008 standards, Canadian Water Resources Journal, 38, 159–167, 2013.

Daigle, A., Caudron, A., Vigier, L., and Pella, H.: Optimization methodology for a river temperature monitoring network for the characterization of fish thermal habitat, Hydrological Sciences Journal, 62, 483–497, 2017.

Datta, S., Karmakar, S., Mezbahuddin, S., Hossain, M. M., Chaudhary, B. S., Hoque, M. E., Abdullah Al Mamun, M., and Baul, T. K.: The limits of watershed delineation: implications of different DEMs, DEM resolutions, and area threshold values, Hydrology Research, 53, 1047–1062, 2022.

Florinsky, I.: Digital terrain analysis in soil science and geology, Academic Press, 2016.

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: How limited training data affects streamflow prediction, Environmental Modelling & Software, 135, 104 926, 2021.

GDAL/OGR contributors: GDAL/OGR Geospatial Data Abstraction software Library, Open Source Geospatial Foundation, https://doi.org/10.5281/zenodo.5884351, 2023.

Geobase: National Hydro Network Data Production Catalogue, Available at: https://www.nrcan.gc.ca/maps-tools-and-publications/survey-plans-and-data/standards-guidelines/10780, accessed: 2023-04-30, 2004.

Gillies, S.: Shapely: Manipulation and analysis of geometric objects, https://github.com/Toblerity/Shapely, 2021.

Gleeson, T.: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity, https://doi.org/10.5683/SP2/DLGXYO, 2018.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R.: Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote sensing of Environment, 202, 18–27, 2017.

Gravelius, H.: Grundrifi der gesamten Gewcisserkunde. Band I: Flufikunde (Compendium of Hydrology, vol. I. Rivers, in German), Goschen, Berlin, 1914.

Gray, M.: Freshwater water atlas user guide, GeoBC Integrated Land Management Bureau. Victoria, BC, 2010.

Guth, P.: Drainage basin morphometry: a global snapshot from the shuttle radar topography mission, Hydrology and Earth System Sciences, 15, 2091–2099, 2011.

Hafen, K. C., Blasch, K. W., Rea, A., Sando, R., and Gessler, P. E.: The influence of climate variability on the accuracy of NHD perennial and nonperennial stream classifications, JAWRA Journal of the American Water Resources Association, 56, 903–916, 2020.

Hafen, K. C., Blasch, K. W., Gessler, P. E., Sando, R., and Rea, A.: Precision of headwater stream permanence estimates from a monthly water balance model in the Pacific Northwest, USA, Water, 14, 895, 2022.

Hershfield, D. M.: On the spacing of raingages, in: Proceedings of the WMO/IASH Symposium on Design of Hydrometerologic Networks, Int. Assoc. Sci. Hydrol. Publ, vol. 67, pp. 72–79, Citeseer, 1965.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., et al.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological sciences journal, 58, 1198–1255, 2013.

Huscroft, J., Gleeson, T., Hartmann, J., and Börker, J.: Compiling and mapping global permeability of the unconsolidated and consolidated Earth: GLobal HYdrogeology MaPS 2.0 (GLHYMPS 2.0), Geophysical Research Letters, 45, 1897–1904, 2018.

Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe, Earth System Science Data, 13, 4529–4565, 2021.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., et al.: Jupyter Notebooks-a publishing format for reproducible computational workflows., Elpub, 2016, 87–90, 2016.

Kovacek, D. and Weijs, S.: British Columbia Ungauged Basins Dataset, https://doi.org/10.5683/SP3/JNKZVT, 2023.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11 344–11 354, 2019.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., et al.: Caravan-A global community dataset for large-sample hydrology, Scientific Data, 10, 61, 2023.

Krause, A., Guestrin, C., Gupta, A., and Kleinberg, J.: Near-optimal sensor placements: Maximizing information while minimizing communication cost, in: Proceedings of the 5th international conference on Information processing in sensor networks, pp. 2–10, 2006.

Latifovic, R., Homer, C., Ressl, R., Pouliot, D., Hossain, S., Colditz Colditz, R., Olthof, I., Giri, C., and Victoria, A.: North American land change monitoring system (NALCMS), Remote sensing of land use and land cover: principles and applications. CRC Press, Boca Raton, 2010.

Lehner, B., Roth, A., Huber, M., Anand, M., Grill, G., Osterkamp, N., Tubbesing, R., Warmedinger, L., and Thieme, M.: HydroSHEDS v2. 0-Refined global river network and catchment delineations from TanDEM-X elevation data, in: EGU General Assembly Conference Abstracts, pp. EGU21–9277, 2021.

Lindsay, J. B.: Whitebox GAT: A case study in geomorphometric analysis, Computers & Geosciences, 95, 75–84, 2016.

Mandelbrot, B.: How long is the coast of Britain? Statistical self-similarity and fractional dimension, science, 156, 636–638, 1967.

Mishra, A. K. and Coulibaly, P.: Hydrometric network evaluation for Canadian watersheds, Journal of Hydrology, 380, 420–437, 2010.

Mutzner, R., Tarolli, P., Sofia, G., Parlange, M. B., and Rinaldo, A.: Field study on drainage densities and rescaled width functions in a high-altitude alpine catchment, Hydrological Processes, 30, 2138–2152, 2016.

PostGIS Project Steering Committee and others: PostGIS, spatial and geographic objects for postgreSQL, https://postgis.net, 2018.

QGIS Development Team: QGIS Geographic Information System, Open Source Geospatial Foundation, http://qgis.org, 2023.

Robinson, N., Regetz, J., and Guralnick, R. P.: EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data, ISPRS Journal of Photogrammetry and Remote Sensing, 87, 57–67, 2014.

Rouhani, S.: Variance reduction analysis, Water Resources Research, 21, 837–846, 1985.

Sassolas-Serrayet, T., Cattin, R., and Ferry, M.: The shape of watersheds, Nature communications, 9, 3791, 2018.

Scholes, R. C., Hageman, K. J., Closs, G. P., Stirling, C. H., Reid, M. R., Gabrielsson, R., and Augspurger, J. M.: Predictors of pesticide concentrations in freshwater trout–The role of life history, Environmental Pollution, 219, 253–261, 2016.

Shavers, E. and Stanislawski, L. V.: Channel cross-section analysis for automated stream head identification, Environmental Modelling & Software, 132, 104 809, 2020.

Tarolli, P. and Dalla Fontana, G.: Hillslope-to-valley transition morphology: New opportunities from high resolution DTMs, Geomorphology, 113, 47–56, 2009.

Thornton, M., Shrestha, R., Wei, Y., Thornton, P., Kao, S., and Wilson, B.: Daymet: Monthly Climate Summaries on a 1-km Grid for North America, Version 4 R1. ORNL DAAC, Oak Ridge, Tennessee, USA, 2022.

Thornton, P. E., Shrestha, R., Thornton, M., Kao, S.-C., Wei, Y., and Wilson, B. E.: Gridded daily weather data for North America with comprehensive uncertainty quantification, Scientific Data, 8, 190, 2021.

U.S. Geological Survey: 1 Arc-second Digital Elevation Models (DEMs) USGS National Map 3D Elevation Program, https://data.usgs.gov/datacatalog/data/USGS:35f9c4d4-b113-4c8d-8691-47c428c29a5b, [Online; accessed 3 March 2022], 2022.

Werstuck, C. and Coulibaly, P.: Hydrometric network design using dual entropy multi-objective optimization in the Ottawa River Basin, Hydrology Research, 48, 1639–1651, 2017.

Werstuck, C. and Coulibaly, P.: Assessing Spatial Scale Effects on Hydrometric Network Design Using Entropy and Multi-objective Methods, JAWRA Journal of the American Water Resources Association, 54, 275–286, 2018.

Wickel, B., Lehner, B., and Sindorf, N.: HydroSHEDS: A global comprehensive hydrographic dataset, in: AGU Fall Meeting Abstracts, vol. 2007, pp. H11H–05, 2007.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific data, 3, 1–9, 2016.

Woodrow, K., Lindsay, J. B., and Berg, A. A.: Evaluating DEM conditioning techniques, elevation source data, and grid resolution for field-scale hydrological parameter extraction, Journal of hydrology, 540, 1022–1029, 2016.

Zhang, J., Condon, L. E., Tran, H., and Maxwell, R. M.: A national topographic dataset for hydrological modeling over contiguous United States, Earth System Science Data Discussions, 2020, 1–26, 2020.

Zhang, W. and Montgomery, D. R.: Digital elevation model grid size, landscape representation, and hydrologic simulations, Water resources research, 30, 1019–1028, 1994.

Table 2. Basin attributes in the BCUB database derived from USGS 3DEP (DEM), NALCMS (land cover), GLHYMPS (soil), and NASA Daymet (climate) datasets.

Group	**Description (BCUB label)**	**Aggregation**	**Units**
Terrain	Drainage Area (drainage_area_km2)	at pour point	km^2
	Elevation (elevation_m)	spatial mean	m above sea level
	Terrain Slope (slope_deg)	spatial mean	° (degrees)
	Terrain Aspect (aspect_deg)	circular mean[2]	° (degrees)
Land Cover[3]	Cropland (land_use_crops_frac_<year>)	spatial mean	% cover
	Forest (land_use_forest_frac_<year>)		
	Grassland (land_grass_forest_frac_<year>)		
	Shrubs (land_use_shrubs_frac_<year>)		
	Snow & Ice (land_use_snow_ice_frac_<year>)		
	Urban (land_use_urban_frac_<year>)		
	Water (land_use_water_frac_<year>)		
	Wetland (land_use_wetland_frac_<year>)		
Soil[4]	Permeability (logk_ice_x100)	geometric mean	m^2
	Std. Dev. Permeability (k_stdev_x100)	geometric mean	m^2
	Porosity (porosity_x100)	spatial mean	% cover
Climate[5]	Annual Precipitation (prcp)	spatial and temporal mean	mm/year
	Daily Minimum Temperature (tmin)		Celsius
	Daily Maximum Temperature (tmax)		Celsius
	Annual Maximum Snow Water Equivalent (swe)		Celsius
	Shortwave Radiation (srad)		W/m^2
	Vapour Pressure (vp)		Pa
	High preciptation frequency (high_prcp_freq)		days/year
	Low preciptation frequency (low_prcp_freq)		days/year
	High preciptation duration (high_prcp_duration)		days
	Low preciptation duration (low_prcp_duration)		days

1. Spatial aspect is expressed in degrees counter-clockwise from the east direction.
2. The <year> suffix specifies the land cover dataset (2010, 2015, or 2020),.
3. Soil parameters follow definitions from Huscroft et al. (2018).
4. Only the climate parameters directly extracted from distinct Daymet source variables are shown here. Additional computed parameters are discussed in subsubsection 2.1.3.
5. A high precipitation event is defined as total daily precipitation greater than 5x the annual mean, and the duration refers to the mean duration of high precipitation events.
6. A low precipitation event is defined as total daily precipitation less than 0.1mm, and the duration refers to the mean duration of low precipitation events.

Table 3. BCUB dataset metadata attributes.

Group	Description (BCUB label)	~~Aggregation~~ Units
Metadata	Region code identifier (region_code)	- ~~-~~
	Pour point[1] (ppt_x, ppt_y)	~~-~~m
	Basin centroid[1] (centroid_x, centroid_y)	~~-~~m
	Soil Flag (soil_flag)	~~-~~binary (0/1)
	Permafrost Flag (permafrost_flag)	~~-~~binary (0/1)
	Geometry Flag (geometry_flag)	binary (0/1)
	Geometry underestimation flag (outside_pct_area_flag)	%
	Geometry overestimation flag (inside_pct_area_flag)	%

1. Geometries are projected to the BC Albers (EPSG:3005) coordinate reference system.

Table 4. Summary of data repository contents.

Filename	Description
BCUB_attributes_20240630.csv	Catchment attributes with geographic coordinates describing the catchment centroid and the outlet (pour point). Catchment polygon geometries are not included for performance.
polygons/*.parquet	Basin attributes and associated catchment boundary, centroid, and pour ppoint geometries are organized into sub-regions to limit file sizes.
BCUB_regions_4326.geojson	Spatial reference file describing the study area sub-regions corresponding to parquet filename prefixes (i.e. VCI_basins.parquet)
MetaData.pdf	General information about the dataset content, formats, versioning, and input data sources.
README.pdf	Basin attribute descriptions and method references.