https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

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

Scott Peckham[1], Keith Jennings[2], Wanru Wu[3], Andy Wood[4,5], and Lauren Bolotin[6]

[1]INSTAAR, University of Colorado, Boulder, CO, USA
[2]Water Resources Institute, University of Vermont, Burlington, VT, USA
[3]Office of Water Prediction, NOAA/NWS, Silver Spring, MD, USA
[4]NSF National Center for Atmospheric Research (NCAR), Boulder, CO, USA
[5]Colorado School of Mines, Golden, CO, USA
[6]Lynker, Boulder, CO, USA

Correspondence: Scott Peckham (scott.peckham@noaa.gov)

Abstract. In the US, several different federal agencies (e.g., the USGS, NOAA, USDA, EPA, and NSF) collect information that has been or continues to be measured for river basins in support of their water-related missions and goals. This information is published online in named data collections, and each data collection has its own set of attributes and objectives. A given basin often has multiple agency IDs and may appear in multiple collections, so there is overlap between them. These collections represent a significant investment of time and money and are a critically important resource for hydrologic modeling and monitoring, whether used operationally or for research. Unfortunately, there is significant heterogeneity across these collections, both in terms of the data they provide but also in terms of how they can be found and effectively accessed. It is also not uncommon for them to contain missing data or errors. Driven by the need to identify the most performant hydrologic model for any given river basin in the US from a collection of available models, the HARBOR project has two key goals. The first is to harmonize and bring together these datasets and associated resources in one place — just as many large cargo ships can be moored in the same harbor — which helps to increase awareness of them while also making it much easier to find, access, and use them. The second is to classify river basins into hydrologically similar groups, since if two river basins are hydrologically similar then it is likely that the same model in a collection will be most performant for both of them. To achieve these goals, a set of Python modules were created, one for each dataset, to augment, clean, and extract information from them. Four different river basin classification methods were applied, given sufficient data, including the Hydrologic Landscape Region (HLR) method, the more process-based Seasonal Water Balance (SWB) method, a simple hydrograph-based method based on modeling with the National Water Model, and the method of using the 12 aggregated ecoregions that were used for the GAGES-II dataset. In order to address shortcomings in the SWB method, we also developed an Extended SWB method and applied it to the 9067 GAGES-II basins in CONUS.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

1 Introduction

How does one identify a river basin, or basins, for hydrologic model evaluation? Ideally, this would be a data-driven selection process, yet researchers often choose the locations with which they are familiar or those that previous studies have used. To mitigate this bias, recent years have seen marked growth in the number of quantitative geospatial datasets that should facilitate more objective basin selections. In the United States, many of these datasets have been created by federal agencies like the US Geological Survey (USGS), the National Oceanic and Atmospheric Administration (NOAA), the Environmental Protection Agency (EPA), and the US Department of Agriculture (USDA). Some have been created in connection with projects funded by the National Science Foundation (NSF) like the Critical Zone Observatories (CZOs), Long-Term Ecological Research (LTER) and National Ecological Observatory Network (NEON). Yet others, usually subsets of the federal agency collections, have been created to support other modeling objectives such as the Model Parameter Estimation Experiment (MOPEX) and the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) datasets. Federal agencies also may create a new collection to supersede or supplement an existing collection. As might be expected, each of these data collections provide a different set of basin attributes, usually in the form of text-based CSV or TSV files (comma or tab separated values), and sometimes also with shapefiles for all of the basins in the collection.

Unfortunately, there is substantial heterogeneity across these basin collections, such as different attributes, different column headings/abbreviations for the same attributes, different measurement units, different methods of organizing files by region, different spatial scales and extents, and missing data. While most datasets are available on publicly accessible websites or through APIs, it can be surprisingly difficult to obtain some of these collections, especially the older ones, which may no longer be available at their original or published online location (e.g., MOPEX). It is also unclear to most potential users how these various basin collections relate to one another, or the extent to which they contain the same basins, and the sea of associated acronyms is daunting. A given basin, associated with a given stream gauge at its outlet, typically has many different agency IDs, such as a USGS 8-to-15-digit ID, a 5-to-8-character NOAA NWS location ID, a GOES (or NESDIS) satellite ID, an 8-to-16-digit Hydrologic Unit Code (HUC), among others.

Such heterogeneity presents several challenges. One, it prevents shared understanding of basin attributes and hydrologic processes in different spatial domains. Two, it hinders efforts to compare model performance across, or even within, basin datasets. And three, it makes it difficult to integrate findings and advances from one dataset to another. This suggests the need to harmonize these disparate datasets into a single, larger dataset that the hydrologic modeling community can deploy. While we designed this work to benefit a wide range of research and operational activities, a key motivating factor was the development of the Next Generation Water Resources Modeling Framework (NextGen) by NOAA (Ogden et al., 2021). The key advancement of NextGen is that it is a plug-and-play modeling framework (vs. a single model) that provides unprecedented modeling flexibility with its ability to use different hydrologic models and process formulations for different catchments or river basins. Notably, the framework enables the unique inclusion and inter-operability of both process-based and data-driven (e.g., machine learning, deep learning) models. This flexibility brings a new challenge, however, in that a desire for local model fidelity may lead to a degree of spatial variation and heterogeneity that is difficult to implement (calibrate), maintain,

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

interpret, and operationalize. Spatial model variation may in some cases provide only marginal performance gains over less heterogeneous selections of models, and regionally consistent modeling patterns may be more acceptable to both forecasters and stakeholders. River basins that are "hydrologically similar" in some appropriate sense may be modeled best by the same hydrologic model. If so, this principle can be used to simplify or automate the matching of the most performant hydrologic models to individual river basins. Moreover, when a model is calibrated for one or more river locations in a class of similar basins, that calibration may be suitable for other, potentially ungauged, members of that class — an approach often described as parameter regionalization.

A comprehensive large domain assessment of model performance to support regionalization objectives requires the ability to test strategies for model selection over a large collection (i.e., sample) of catchments spanning the entire modeling domain. This observation motivates the work described in this paper, which has two primary goals. The first goal is to collect and examine many different river basin data collections, to simplify access to them, to some extent harmonize and collate them, and to see what attributes they provide that could be used for river basin classification. The second goal was to examine different river basin classification schemes (and their data requirements) as a means of placing river basins in a manageable number of classes with other hydrologically similar basins. For this goal we seek a river basin classification method that identifies which hydrologic processes, climate variables, and basin attributes are most important for distinguishing the degree to which two river basins are hydrologically similar.

To accomplish these goals, we acquired a diverse collection of basin datasets and then developed a set of Python utilities, with a subset tailored to each dataset, to collate, clean, augment, and extract information from them. We describe the river basin datasets in detail in the subsequent section: Overview of Existing River Basin Data Collections. We then describe the set of utilities in the following section titled: Python Utilities for Working with River Basin Data Collections. The organizational structure of this new basin repository is described in: Organization of HARBOR.

To fulfill the second key goal of this work, several river basin classification systems were examined, and these are summarized in the section titled: Existing River Basin Classification Systems. One of these systems, termed the Seasonal Water Balance method, is relatively more process-based and seems particularly promising after work to resolve initial shortcomings. This new version is described in the section titled: An Extended Seasonal Water Balance Classification Method.

2 Overview of Existing River Basin Data Collections

In this section, brief descriptions are provided (with references) for many existing river basin data collections. Figure 1 shows the extent to which these collections overlap with one another. The function *compare_basin_ids()* in the Python module *data_utils.py* was used to create this diagram.

CAMELS (Catchment Attributes and Meteorology for Large-sample Studies). CAMELS contains 671 CONUS watersheds with minimal human impact that span a wide range of hydroclimatic conditions, originally providing associated datasets from 1980 to 2015 (Newman et al., 2015; Addor et al., 2017; NCAR, 2024) All of the CAMELS basins are contained in the

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

set of GAGES-II SB3 basins, which is a subset of the GAGES-II Reference basins (see below). 52 CAMELS basins are also in the MOPEX dataset described below. Notably, CAMELS paired catchment meteorological time series from several common forcing datasets, observed streamflow time series, modeling results from a common hydrology model, and an unusually large collection of catchment geophysical attributes, enabling it to become the foundational dataset for the current rise of machine learning for streamflow simulation, and a community-wide hydrology benchmarking dataset (Newman et al., 2015).

Caravan. An open community dataset of meteorological forcing data, catchment attributes, and discharge data for catchments around the world. Inspired by CAMELS, multiple countries developed CAMELS dataset extensions for their own countries, which Kratzert et al. (2023) collected and extended, including national datasets such as the CAMELS-CL dataset for Chile (Alvarez-Garreton et al., 2018).

HYSETS. This is a multisource dataset with information for 14,425 basins in North America. See Arsenault et al. (2020) and Mai et al. (2022). It includes a wide array of hydrometeorological data required to perform hydrological and climate change impact studies, namely (1) watershed properties including boundaries, area, elevation slope, land use and other physiographic information; (2) hydrometric gauging station discharge time-series; (3) precipitation, maximum and minimum daily air temperature time-series from weather station records and from (4) the SCDNA infilled gauge meteorological dataset; (5) the NRCan and Livneh gridded interpolated products' meteorological data; (6) ERA5 and ERA5-Land reanalysis data; and (7) the SNODAS and ERA5-Land snow water equivalent estimates. Watersheds with regulated (non-natural) flows were excluded, thus only a subset of USGS gauging station data is included. All data is available online at: Arsenault et al. (2011). Watershed properties are listed in the file: HYSETS_watershed_properties.txt and include the latitude and longitude of the station/outlet in the last 2 columns. This dataset is not included in the collated basin TSV file.

MOPEX (Model Parameter Estimation Experiment). This dataset has information for 431 well-monitored, lower-impact basins with a focus on parameter estimation for hydrologic models. Starting from a set of 1861 potential MOPEX basins, only 431 were considered to have a sufficient areal density of rain gauges. Time series data spans 1948 to 2003 for basins mostly in the eastern half of CONUS, with a minimum record length of 10 years and often more than 20 years. See Schaake et al. (2006) and Duan et al. (2006). MOPEX was an international project that hosted at least 5 workshops in different countries between 1999 and 2005, with a focus on high-quality data sets for unregulated, intermediate-sized basins (500 to 10,000 km^2). MOPEX was augmented with many additional attributes as part of a 2009 Hydrologic Synthesis Project (Durcik et al., 2009; Troch et al., 2018). The paper by Berghuijs et al. (2014) used data for 321 MOPEX basins to develop the Seasonal Water Balance (SWB) basin classification method, discussed in a subsequent section. Brooks et al. (2011) also used data for a subset of MOPEX basins.

NOAA-NWS River Forecast Center (RFC) Basins. The USA is divided into 13 RFC regions, each with forecasts (and data) for many rivers of various sizes within that region (Figure 3). Each RFC provides data on a website, with a URL of the form:

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

https://www.weather.gov/wgrfc/. See NOAA (2024b). This dataset includes a shapefile called *ba12my15.shp* that contains 9370 basins. The USGS and NOAA work together to monitor over 9109 basin DCPs (Data Collection Platforms) that upload data to a geostationary GOES satellite. These DCPs/gages then have both a USGS Site ID and an NWS Location ID (NWSLID). Many of these are part of the HADS network of DCPs. There is a NOAA HADS-to-USGS crosswalk that maps many USGS Site IDs to corresponding NWS Location IDs. The SERFC is responsible for Puerto Rico. In the following subsections we document a few additional characteristics of these RFC data sets.

NWS Location IDs and Runoff Zone IDs for MBRFC. The Missouri Basin RFC (MBRFC), one of the 13 NOAA River Forecast Centers, subdivides basins into multiple runoff zones for modeling purposes. While the basins themselves have an associated NWS location ID that follows the 5-character NWSLID scheme, the IDs used for the runoff zones that comprise a basin depend on whether or not the basin is mountainous. For runoff zones in mountainous regions, which are subdivided by elevation, IDs are constructed by appending extra letters like *LWR* or *UPR* to the NWSLID. For runoff zones in non-mountainous regions, a unique 3- or 4-digit numerical ID is used, ranging from 101 to 3297. Shapefiles may contain boundaries for both the basins themselves, as well as the runoff zones. (Note: MBRFC employees kindly provided us with a mapping of runoff zones to their corresponding basins. The Missouri Basin is also divided into 20 regional basins called forecast groups.) Note that the outlet of the most downstream runoff zone in a given basin will coincide with that basin's outlet, so their outlet latitude and longitude will be the same. Other RFCs (e.g., ABRFC, CBRFC, NCRFC, NERFC) also add additional letters after the 5 standard alphanumeric characters as a means of grouping nearby sites or discriminating model zones within a gage's drainage area. Elevation zones are common in NWS watershed modeling, and are usually indicated by characters such as L, M, and U to denote lower, middle, and upper zones.

Incomplete Metadata for Some Basins.
Four of the dataset folders that start with **NOAA_** contain alternate datasets for the basins associated with the 13 NOAA RFCs. For example, one folder has info from a beta version of a new API (note: official release is now available), and one has info from a NOAA HADS-to-USGS crosswalk for gages in the HADS/GOES network. Some utilities in *rfc_utils.py* attempt to utilize these different datasets to fill in missing information. For example, some datasets may not provide the RFC, WFO/CWA, or HSA codes for all sites. See Figure 4 for a map of the WFO/CWA regions.

NSF CZO (Critical Zone Observatories). In 2006, the U.S. National Science Foundation (NSF) sent out the first Critical Zone Observatory (CZO) solicitation (NSF06-588) that was followed by another solicitation in 2012 (NSF12-575). The CZO program funded 10 CZOs spanning different climatic and physiographic environments that operate as "environmental laboratories". (However, the Christina CZO is no longer funded.) In 2019, the NSF sent out a solicitation for a Critical Zone Collaborative Network (CZ-Net) to support "thematic clusters" and a "coordinating hub". The term "critical zone" refers to the permeable, near-surface layer of the Earth that extends from bedrock to treetops and is "where rock, soil, water, air, and living organisms interact". Research at CZOs focuses on interconnected, catchment-scale processes (e.g., chemical, biological, phys-

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

ical). See Brantley et al. (2017), NSF-CZO (2021), and Wlostowski et al. (2021). The last paper includes two supplementary data files, including a spreadsheet with metadata for many of the CZO watersheds. UNWI (2019) is a HydroShare dataset that identifies the 22 USGS gages that are adjacent to a CZO watershed. Since CZO watersheds have been intensively studied and often have good observational data, they have been included in the repo and are indicated in the *Is_CZO* column of the collated TSV file.

NSF LTER (Long-Term Ecological Research). The NSF-funded LTER program began in 1980 and focuses on ecological research with long-term observations at 30 sites across the US (27 still active). Many of the LTER sites contain river basins and the closest USGS gage and distance to it (in km) is indicated in the *Closest_Site_ID* and *Closest_Site_Dist* columns of the collated TSV file. There are 22 LTER site centroids within 10 km of a USGS gage. See Franklin et al. (1990), Hobbie et al. (2008), Gosz et al. (2010), and Müller et al. (2010).

NSF NEON (National Ecological Observatory Network). The NEON project began in 2011 with partial funding from the NSF and consists of 81 field sites across US with >30 years of ecological and climatology data. A CSV file with NEON Field Site metadata can be downloaded at NEON (2024b). Field sites may be aquatic (rivers or lakes) or terrestrial. The 33 NEON watersheds have non-empty entries in the columns: *field_watershed_name* and *field_watershed_size_km2*. Latitude and longitude entries are presumed to be for watershed outlets but could be centroids **CHECK THIS**. The closest USGS gage and distance to it (in km) is indicated in the *Closest_Site_ID* and *Closest_Site_Dist* columns of the collated TSV file, and all are within 22 km. See NEON (2024a, c). There is also an R package for data access; see NEON (2024d).

USDA/ARS Experimental Watershed Network. A network of 771 watersheds with a focus on agriculture and soil erosion, many with long discharge records. Several of these have been the subject of many studies and are well-known to hydrologists, such as: Goodwater Creek (MO), Goodwin Creek (MS), Little River (GA), Little Washita (OK), Lucky Hills (AZ), Pigeon Roost (MS), Reynolds Creek (ID), Upper Sheep Creek (ID), and Walnut Gulch (AZ). See Goodrich et al. (1994, 2011, 2016, 2020) and Cook (2021). Data for these basins is in the STEWARDS database. See USDA (2024) for watershed metadata, runoff data, rain gauge data and maps.

USDA SCAN Sites. A set of 212 stations measuring soil moisture content at several depths, air temperature, relative humidity, solar radiation, wind speed and direction, liquid precipitation, and barometric pressure. See USDA-NWCS (2024a, b).

USDA SNOTEL (SNOwpack TELemetry Network). Snow data for 899 sites across 11 US states. See USDA-NWCS (2024c, d).

USGS FPS (Federal Priority Streamgages). A set of 4756 monitoring stations and the "backbone" of the larger USGS Streamgaging Network. "These sites are eligible for federal funding, as available. As of January 16, 2024, 3,491 of these sites

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

were active, collecting real time streamflow and/or water level data to meet federal needs." See Normand (2021) and USGS-FPS (2024). See Figure 2.

USGS GAGES-II (Geospatial Attributes of Gages for Evaluating Streamflow, version II). This widely-used data set provides numerous attributes for 2057 *Reference* (least-disturbed) sites and 7265 *Non-reference* basins (9322 total). As of 2009, 1,633 of the *Reference* sites had 20+ years of record since 1950. It has all CAMELS basins and all but 7 MOPEX basins. See Falcone et al. (2010), Falcone (2011) and Falcone (2017). This dataset also includes shapefiles for all of the Reference and Non-Reference basins. A subset of the basins in GAGES-II are identified with "yes" in a column with the heading *HCDN-2009* in the spreadsheet *gages2_sept30_2011_conterm.xlsx*. These 743 basins were flagged as potentially useful for a future hydro-climatic study similar to the original HCDN study because they fulfill all of the following criteria:

1. have 20 years of complete and continuous flow record in the last 20 years (water years 1990-2009), and were thus also currently active as of 2009;
2. are identified as being in current Reference condition according to the GAGES-II classification;
3. have less than 5 percent imperviousness as measured from the 2006 National Land Cover Database; and
4. were not eliminated by a review from participating state Water Science Center evaluators.

USGS GAGES-II CONUS SB3. This data set provides many additional attributes for all GAGES-II Reference basins that lie within the conterminous US (CONUS). It therefore excludes 67 sites in Alaska, 27 in Hawaii, and 16 in Puerto Rico, resulting in extended data for 1947 (i.e. 2057 - 110) basins. See *Child Item 3* in the Russell et al. (2018) dataset. Since many of the filenames in this dataset begin with SB3 (where SB stands for Selected Basins), we are using SB3 to refer to this dataset. It includes all CAMELS basins, 88 MOPEX basins, and 561 HCDN basins. It was developed to investigate refinements of a regression-based method for prediction of flow-duration curves for selected streamgages in CONUS (Over et al., 2018). An error was discovered and fixed in this dataset, where 16 of the USGS site IDs in the file *SB3_untransfBCs.new.xlsx* were missing a leading "0".

USGS HCDN (Hydro-Climatic Data Network). Slack and Landwehr (1992a) describe HCDN as: "A national data set of streamflow records that are relatively free of confounding anthropogenic influences has been developed for the purpose of studying the variation in surface-water conditions throughout the United States." This dataset contains data from WATSTORE for USGS basins with relatively long streamflow records and was applied to the study of climate fluctuations. While it is said to provide data for 1659 basins, it actually provides data for 1703 basins. Basins are in CONUS, Alaska, Hawaii, Puerto Rico, and the US Virgin Islands. Note that 1523 HCDN basins are included in GAGES-II and 595 are Reference basins. See Slack and Landwehr (1992a, b) and Slack et al. (1994). Falcone (2011) noted that the HCDN "identifies stream gages which at some point in their history had periods which represented natural flow, and the years in which those natural flows occurred were identified."

USGS HLR (Hydrologic Landscape Regions). The concept of hydrologic landscapes was introduced by Winter (2001). Soon after, Wolock et al. (2004) developed HLR into a basin classification system (described in detail in a later section) and applied

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

it to the entire US using a GTOPO30 HYDRO1K DEM for North America and GIS (e.g., basin delineation) methods. Note that HYDRO1K uses a Lambert Azimuthal Equal Area projection. The HLR system groups watersheds into 20 classes (or noncontiguous regions) on the basis of similarities in land-surface form, geologic texture (soil and bedrock types), and climate characteristics. Valid HLR codes range from 1 to 20, but a value of 0 is used to indicate an unclassified basin in the vector-format version. Wolock et al. (2004) assigned HLR codes to 43,931 small (average size of 212 square km) watersheds in the 50 United States. This set of watersheds spans the entire US so that an HLR code is assigned to every square kilometer. The GIS files generated by this work are available online (Wolock, 2003) in both vector and raster formats, but the vector data (shapefile, etc.) appears to be a draft version and suffers from various issues. (These are documented in a README file in the repo.) Blackburn-Lynch et al. (2017) developed a separate set of hydraulic and bankfull geometry parameters for each of the 20 hydrologic landscape regions (HLRs) and the results showed marked improvement over other region-based approaches. Blackburn-Lynch et al. (2017) also used data for 2,856 independent basins across CONUS.

USGS NWIS Web Data. USGS water data is made available through the National Water Information System (NWIS), which has a web interface. Real time (current) data can be downloaded from: USGS-NWIS (2024a). Daily values can be downloaded from: USGS-NWIS (2024b). The complete inventory of site data is available at: USGS-NWIS (2024d), but has a "maximum retrieval size" that restricts the amount of data that can be downloaded in a given search. Step-by-step instructions for how to download data from these URLs is given in "how to" text files in the *USGS_NWIS_Web/Data* folder of the HARBOR repo. HARBOR contains separate TSV files for sites of type "Stream" from the NWIS *current*, *daily*, and *inventory* (i.e. all) collections. These are in the *USGS_NWIS_Web/Data* folder, with names:

Current: *NWIS_Stream_Sites_Current.tsv* (11,215 sites)

Daily: *NWIS_Stream_Sites_Daily.tsv* (27,915 sites)

Inventory: *NWIS_Stream_Sites_All.tsv* (159,375 sites).

We discovered that 266 sites occur in the NWIS Daily dataset that are not included in the NWIS Inventory dataset, even though the latter is supposed to be all-inclusive. It turned out that these sites were located outside of the US, and included sites in Canada, Okinawa, and Afghanistan. Note that the USGS GAGES-II dataset is also a subset of the USGS NWIS basins, consisting of 9322 basins, of which 2057 are labeled as "Reference" (vs. "Non-reference") basins. All CAMELS basins (671) and all MOPEX basins (471) are also a subset of the set of all USGS NWIS basins of type *Stream*. All of the CAMELS basins and all but one of the MOPEX basins are also a subset of the set of *daily value* USGS NWIS basins of type *Stream*. (The USGS site ID of that one basin is 13298500, with site name *Salmon River near Challis, ID*, and that basin is in both the HCDN and GAGES-II Non-Reference data sets.)

USGS Streamgaging Network. Contains a subset of USGS NWIS Web basins (about 11,340). See Normand (2021, Figure 1), USGS (2021), and Eberts et al. (2018).

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

USGS National Streamflow Network. This is a subset of the "USGS Streamgaging Network" (about 8,460) and the USGS Federal Priority Streamgages are a subset of this network (about 3470). See Normand (2021, Figure 1).

3 Organization of HARBOR

In building HARBOR, we were guided by the FAIR (Findable, Accessible, Interoperable, and Reproducible) dataset principles that have been made popular by Wilkinson et al. (2016). To this end, (1) all of the Python utilities used to create the Repo are open-source, well-documented, and available in a GitHub code repository, (2) all required GIS work was performed with an open-source GIS application called QGIS and all QGIS workflows are described step-by-step in README files, (3) a URL for each dataset downloaded and used is provided, (4) a uniform organizational structure was created for the Repo, (5) any data oddities discovered were documented, (6) other helpful resources associated with a dataset, such as documentation, APIs, and URLs are included with the Repo, and (7) an extensive set of references (with DOIs) is included with this paper.

Most of the folders in the repo are named to reflect a specific dataset, often associated with a particular federal agency. For example, several folder names start with a federal agency name like: **NOAA_**, **NSF_**, **USDA_**, or **USGS_**. Within a dataset folder, there will usually be a folder called **Data**. The Data folder will contain a Mac-based shortcut (.webloc) to the website that the dataset was downloaded from, starting with a double underscore, "__". Often it will also contain a file called **__README.txt** with helpful information specific to that dataset. If the dataset is not too large it is also included in this Data folder (possibly zipped), or at least key portions of it. Files uploaded by browser to GitHub have a limit of 25 MB, while files uploaded by command line can be up to 100 MB. The dataset folder will usually also contain a folder called **_New** that contains files generated from the datasets by the set of Python utilities or by **QGIS**. In many cases, the dataset folder will also contain folders called **Docs** (with additional documentation relating to the dataset), **Papers** (with PDF files for key papers that describe or use the dataset), and **URLs** (with additional Mac-based shortcuts to related websites). If there is a GitHub repo associated with the dataset, there may also be a folder called **GitHub** with a link to that repo. In addition to the dataset folders, there are a few other folders such as:

__Collated (with combined, selected information for many of the datasets; see Table 1.)

__Docs (with general docs that apply across multiple datasets)

APIs_or_Services (with links to websites that provide an API or service)

SWB (with information about the Seasonal Water Balance basin classification method).

If you intend to use the Python utilities to re-create or augment files in the **_New** folders, you should first check the **_New** folder for each dataset and unzip any ".zip" files you find there. These will typically be Python dictionaries or datasets that speed up computation, saved into .pkl (pickle) or .npy (numpy) files. Not all of the datasets with a folder in the repo have been merged into the final, collated TSV file, in the **__Collated** folder. Some have only been included for reference, such as the **Caravan** and **HYSETS** dataset folders.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Attribute	Description
Site_ID	USGS site ID, 8 to 15 digits
NWS_Loc_ID	NOAA NWS location ID, 5 to 8 alphanumeric characters
GOES_ID	GOES satellite ID for a DCP (data collection platform) assigned by NESDIS
RFC	5-letter abbreviation for a NOAA River Forecast Center
WFO/CWA	3-letter abbreviation for Weather Forecast Office or County Warning Area.
HSA	NOAA NWS Hydrologic Service Area
HUC	USGS Hydrologic Unit Code (in most cases 12 digits; sometimes 8)
Site_Name	Original USGS site name, often with many abbreviations
Site_Type	The USGS site type (e.g., Stream, Lake, Atmosphere, Well, etc.)
Stage_Data	c=continuous, i=intermittent, followed by A for active, I for inactive, or N for never recorded
PEDTS_Obs	A 5-character parameter code that describes what quantity was observed or predicted at a location. PE=Physical Element, D=Duration Code, T=Type Code, S=Source Code. See NOAA (2012).
State_Code	2-letter US state code
Country_Code	country code
Outlet_Lon	longitude of basin/gage outlet
Outlet_Lat	latitude of basin/gage outlet
Outlet_Elev	elevation of basin/gage outlet
Elev_Units	elevation units
Area	drainage area above basin/gage outlet
Area_Units	area units
Horiz_Datum	horizontal datum
Vert_Datum	vertical datum
Minlon	westernmost longitude of geographic bounding box
Maxlon	easternmost longitude of geographic bounding box
Minlat	southernmost longitude of geographic bounding box
Maxlat	northernmost longitude of geographic bounding box
Long_Name	expanded USGS site name, without abbreviations
Closest_Site_ID	the closest USGS site ID
Closest_Site_Dist	distance to the closest USGS site ID
Site_URL	URL associated with the USGS site ID
HUC_URL	URL associated with basin's Hydrologic Unit Code
NWS_URL	URL associated with NWS Location ID
Status_as_FPS	status of the gage (active or inactive), according to FPS
Start_Date	start date for data collection, if known
End_Date	end date for data collection, if known
Eco_Region	GAGES-II aggregated ecological region name (of 12)
HLR_Code_Outlet	USGS Hydrologic Landscape Region code (0 to 20) at outlet
SWB_Class	Extended Seasonal Water Balance class (of 10 classes), if computable
Hgraph_Type	Hydrograph type from NWM 3 work, if known
Aridity	Budyko aridity index, (mean annual PET / mean annual precipitation)

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Snow_Fraction	Fraction of precipitation that falls as snow
Is_USGS_NWIS_Web	Is site in dataset from official NWIS website?
Is_GAGES2_Any	Is site in the USGS GAGES-II dataset?
Is_GAGES2_Ref	Is site in the USGS GAGES-II Reference dataset?
Is_GAGES2_SB3	Is site in the USGS GAGES-II CONUS SB3 dataset?
Is_FPS	Is site in Federal Priority Streamgage dataset?
Is_HCDN	Is site in Hydro Climatic Data Network dataset?
Is_RFC	Is site in NOAA River Forecast Center dataset?
Is_CAMELS	Is site in the CAMELS dataset?
Is_MOPEX	Is site in the MOPEX dataset?
Is_CZO	Is site in the NSF CZO dataset?
Is_LTER	Is site in the NSF LTER dataset?
Is_NEON	Is site in the NSF NEON dataset?
Is_ARS	Is site in the USDA ARS experimental watershed dataset?

Table 1: List of 53 attributes included in the collated TSV file of HARBOR.

4 Hydrologic Similarity

If two river basins have some appropriate set of geophysical attributes in common (including climate attributes), one may expect that they will be "hydrologically similar" in terms of their streamflow response. Available attributes for classifying basins will depend on whether basins are gauged or ungauged. For example, **hydrologic signatures** can only be computed when observational, time series data (e.g., discharge, precipitation rate, or evaporation rate) are available. Examples include:

Aridity index: There are multiple definitions that attempt to measure aridity or dryness. (One definition: Similar to evaporation ratio but using potential vs. actual evaporation in the numerator. The Budyko curve (Budyko, 1958) plots the evaporation ratio vs. this aridity index.)

Evaporation ratio: The total (actual) volume of water lost to evaporation from a given watershed (integrated over some period of time), divided by the total (liquid-equivalent) volume of precipitation that falls on that watershed during that time.

Runoff ratio: The total volume of water that flows out of a given watershed (integrated over some period of time), divided by the total (liquid-equivalent) volume of precipitation that falls on that watershed during that time. Lower values indicate a greater loss of water due to processes such as infiltration, evaporation, and transpiration.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Snowfall fraction: The total volume of liquid-equivalent precipitation that falls as snow (integrated over some period of time), divided by the total volume of liquid-equivalent precipitation. (Snow runoff fraction is a closely related concept.)

5 Existing River Basin Classification Systems

Different agencies and groups have developed categorical classifications for watersheds to support varying objectives. For clarity, the entire phrase *river basin classification system* is used here since *basin classification system* is also associated with sedimentary basins in a geologic context.

5.1 Hydrologic Landscape Regions (HLR)

The dataset used by the USGS in defining HLRs consists of 43,931 disjoint watersheds that span the entire United States (all 50 states), each roughly 200 square kilometers in size. The descriptions of the 20 Hydrologic Landscape Regions are based on a relatively small set of tags or "descriptors". See Winter (2001), Wolock (2003, Table2), and Wolock et al. (2004). These include a climate descriptor (arid, semiarid, subhumid, humid, or very humid), followed by a landscape (or land-surface form) descriptor (plains, plateaus, playas, or mountains) followed by a simple soil descriptor (permeable or impermeable), followed by a bedrock descriptor (permeable, impermeable, or unspecified). Note that the HLR method does not require a time series of observations for classification. See Table 2 for descriptions or the 20 HLR classes and their corresponding "HLR numbers". See Figure 5 for a hillshaded map of the HLR regions in CONUS.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Class	Description
0	Unclassified
1	Subhumid plains with permeable soils and bedrock
2	Humid plains with permeable soils and bedrock
3	Subhumid plains with impermeable soils and permeable bedrock
4	Humid plains with permeable soils and bedrock
5	Arid plains with permeable soils and bedrock
6	Subhumid plains with impermeable soils and bedrock
7	Humid plains with permeable soils and impermeable bedrock
8	Semiarid plains with impermeable soils and bedrock
9	Humid plateaus with impermeable soils and permeable bedrock
10	Arid plateaus with impermeable soils and permeable bedrock
11	Humid plateaus with impermeable soils and bedrock
12	Semiarid plateaus with permeable soils and impermeable bedrock
13	Semiarid plateaus with impermeable soils and bedrock
14	Arid playas with permeable soils and bedrock
15	Semiarid mountains with impermeable soils and permeable bedrock
16	Humid mountains with permeable soils and impermeable bedrock
17	Semiarid mountains with impermeable soils and bedrock
18	Semiarid mountains with permeable soils and impermeable bedrock
19	Very humid mountains with permeable soils and impermeable bedrock
20	Humid mountains with permeable soils and impermeable bedrock

Table 2: Descriptions of the 20 Hydrologic Landscape Region (HLR) classes, from Wolock et al. (2004).

The 43,931 USGS HLR basins have an average drainage area of 200 square kilometers, while the low order basins that can each be modeled individually in NextGen (by a different model) are smaller, ranging in size from 3 to 15 km^2. A separate flow routing algorithm (t-route), based on a Muskingum-Cunge kinematic wave formulation, is used to combine the streamflow from low order basins to provide predictions for larger basins (reference). Since the NextGen "low order basins" (or smallest Hydrofabric basins) are relatively smaller, they tend to have a well-defined HLR number. Those contained within one of the USGS HLR basins presumably inherit the same HLR number as the enclosing basin. Wolock et al. (2004, Table 8) shows that the HLR system partially captures and/or allows one to infer the "primary hydrologic flow path", given the 3 options of overland flow, shallow groundwater, and deep groundwater. This appears to be because it considers the permeability of both soil and bedrock, which has a direct bearing on the infiltration process. Different hydrologic models tend to perform better for one of these 3 hydrologic flow paths, which could facilitate pairing models to HLR numbers. Similarly, the climate descriptors (arid, semiarid, humid, subhumid, etc.) to some extent capture the relationship between precipitation and potential evaporation. In summary, this system at least partially reflects hydrologic process dominance, with the possible and important exception of

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

the snow melt process.

5.2 Seasonal Water Balance Classification (Berghuijs et al., 2014)

Berghuijs et al. (2014) have proposed a basin classification system based on "seasonal water balance" (SWB) similarity. This system has 10 classes or clusters, as determined from applying a lumped, conceptual rainfall-runoff model, "FLEX_l" (Berghuijs et al., 2014, see Figure 1 and Table 1) to 321 MOPEX basins across CONUS, ranging from 67 to 10,329 sq km, with limited human influence. (This work starts with 372 MOPEX basins, but 51 are removed due to an unacceptable model fit.) The FLEX_l model consists of 5 coupled stores, namely: snow (CR), vegetation interception (IR), unsaturated zone (UR), saturated groundwater (SR) and fast runoff (FR). Its equations and parameters are summarized in Berghuijs et al. (2014) Table 1, and the model is calibrated using the MOSCEM-UA algorithm with 10,000 iterations. Best-fit model results are used to divide the basins into 10 clusters with similar properties based on seven components of the mean seasonal water balance, namely: Pn (precipitation), Q (streamflow), Ps (snowmelt), Ss (snow storage), Ea (evaporation), Su (storage), and D (deficit). See Figure 6 in Berghuijs et al. (2014) . The properties of the 10 SWB classes are summarized in Table 3 of that paper, which is partly reproduced here in Table 3. The resulting process-based similarity framework is ultimately based on three hydroclimatic indices that represent (1) an aridity index (ratio of annual potential evaporation to annual precipitation), (2) seasonality and timing of precipitation (including the extent to which precipitation and potential evaporation are in or out of phase), and (3) snowiness (fraction of precipitation falling as snow). Note that, unlike HLR, this system explicitly includes the snowmelt process and classes B1 and B2 each have a large snowmelt component. We extended the SWB classification system and applied it to the 9067 GAGES-II CONUS basins as described in a subsequent section. Note that Berghuijs et al. (2014, p. 5640) build on prior basin classification work by Kennard et al. (2010), Sawicz et al. (2011), Coopersmith et al. (2012), and Ye et al. (2012).

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Class	Description
A1	Humid catchments where precipitation and evaporation are out of phase. Consequently, large soil water and streamflow variations occur and streamflow is perennial. Vegetation: coniferous.
A2	Semiarid catchments where precipitation and evaporation are out of phase. Consequently, large soil water and streamflow variations occur and streamflow can be perennial or intermittent. Vegetation: coniferous/shrubs.
A3	Arid catchments where precipitation and temperature are out of phase. Consequently, soil water and streamflow variations occur and streamflow can be intermittent. Vegetation: shrubs.
B1	Mountainous humid catchments where snow storage causes a delay in the streamflow and soil water recharge peak. Catchments have perennial streamflow. Vegetation: coniferous.
B2	Mountainous semiarid catchments where snow storage causes a delay in the streamflow and soil water recharge peak. Catchments have perennial streamflow. Vegetation: coniferous.
C1	Semiarid catchments where precipitation and evaporation are in phase. Streamflow and storage variations of both soil water and snow are small. Streams may fall dry but can be perennial. Vegetation: Some short grass prairie, but mainly long grass prairie.
C2	Arid catchments where precipitation and evaporation are in phase. Seasonal streamflow and storage variations of both soil water and snow are very small. Streams are intermittent. Vegetation: short grass prairie.
D1	Humid catchments where precipitation and evaporation are slightly out of phase. Catchments have soil water storage variations and a slightly seasonal streamflow regime with low flows during summer. Vegetation: mixed deciduous and coniferous.
D2	Humid catchments where precipitation and evaporation are slightly in phase. Catchments have small soil water storage variations and a fairly constant seasonal streamflow regime. Vegetation: deciduous.
D3	Humid catchments where precipitation and evaporation are slightly in phase. Catchments have soil water and snow storage variations with a soil water and streamflow increase in spring. Vegetation: deciduous.

Table 3: Descriptions of the 10 Seasonal Water Balance classes, from Berghuijs et al. (2014).

5.3 Hydrograph-based Classifications

Basins can also be classified based on the shapes of their observed or modeled hydrographs into groups such as: flashy, slow, snow-dominated, low-flow, regulated, and "normal". These six classes were used in a National Water Model (NWM) v3 (Cosgrove et al., 2024) analysis which used a procedure based on visual inspection to classify USGS basins with mean streamflow greater than a very small threshold. These classifications have been included in the combined TSV file when available. Berghuijs et al. (2014, Figure 6) showed that basins in the same SWB class also tend to have hydrographs with very similar shapes.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

5.4 Data-driven Machine-Learning Methods

Using watersheds in the CAMELS dataset and their associated attributes, Bolotin et al. (2022) showed how a machine-learning algorithm (Random Forest) could be used to match basins to the most performant of 3 models in NextGen, with an R^2 value of 0.75-0.80. This work considered the process-based Conceptual Functional Equivalent (CFE) model, a Long Short-Term Memory (LSTM) machine-learning model, and the National Water Model version 2.0. Performance was evaluated with the Normalized Nash Sutcliffe Efficiency. It was found that certain models dominated in different regions across CONUS. Further analysis could elucidate the relationship between patterns of model performance and different methods of hydrologic classification.

5.5 Hydrologic Signatures

Hydrologic signatures are metrics calculated from streamflow time series, and sometimes additionally precipitation, soil moisture, or evapotranspiration time series, that characterize the hydrologic regime of a watershed (Gupta et al., 2008; McMillan, 2020a, b). Consequently, watersheds can be classified by their hydrologic signatures, indicating the dominant processes across space, or even across seasons within singular watersheds (Gnann et al., 2020; Wu et al., 2021). The CAMELS dataset includes hydrologic signatures such as baseflow index, runoff ratio, and slope of the flow duration curve for the watersheds included in the collection (Addor et al., 2017). River basin collections with catchment attributes have been used in various applications to predict hydrologic signatures using machine learning, elucidating the link between catchment and climate characteristics and hydrologic processes (Wu et al., 2021; Addor et al., 2018; Bolotin and McMillan, 2024). Our collation of available basin collection datasets will facilitate future large sample studies leveraging hydrologic signatures by easing the selection of study watersheds and identification of available data associated with these watersheds.

6 Extended Seasonal Water Balance Classification

Recall that each of the 10 classes in the Seasonal Water Balance (SWB) method is ultimately defined using a range of parameter values for 3 indices, namely **aridity**, **seasonality** and **snowiness**. Therefore, each class corresponds to a rectangular box or "cuboid" in a 3-dimensional space, and these boxes may touch but may not overlap. When we applied the original SWB method proposed by Berghuijs et al. (2014) to the GAGES-II CONUS dataset (which has 9067 basins), we found that 3971 basins did not get classified.

In order to understand and resolve the issue, we used a graphical capability in Mathematica (Wolfram Research Inc., 2024) that allowed us to plot and label the 10 3D "cuboids" as semi-transparent 3D boxes with different colors in a plot that that could be rotated with the computer mouse and viewed from different angles. This resulted in Figures 6 and 7 which clearly show gaps or unclassified regions of parameter space and explains why many basins were left unclassified. Keeping in mind that the 3D cuboids associated with different classes cannot overlap, each cuboid was expanded along each axis by the minimum amount necessary to span the entire 3D parameter space. In most cases, there was no flexibility or ambiguity about how much each

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

3D box had to be expanded due to spatial constraints. However, Figure 7 shows a gap along the seasonality or δ_p axis between classes A1, A2 and A3 an classes C1, C2, D2, and D3. In this case, the gap was closed by expanding the cuboids to meet at a value of $\delta_p = -0.1$. All things considered, this seems to be the best value, but possible values lie in $[-0.4, -0.1]$. The gap between C1, C2 and B1, B2 was closed by increasing the upper bound on f_s for classes C1 and C2. Also, it made sense to add an 11th class, B3, by direct analogy to class A3, but none of the GAGES-II CONUS basins fell into this new class. Class B3 stacks on top of B2 just as A3 stacks on top of A2. Class D1 (not shown) is unusual because it has only a single value of zero for f_s and is therefore a 2D rectangle vs. a 3D cuboid. The cuboids of our extensions to the SWB method are shown in Figure 8 and span the entire parameter space.

As a side note, the range of parameters used to define the 10 classes in the Berghuijs et al. (2014) paper are given in both Figure 7 and Table 3 of that paper, but those parameter ranges do not always agree. We took the broadest of the two ranges given as the official definition of the original SWB classes. The range of values used to define each class, for both the original and extended SWB methods, are given in the Python module *swb_utils.py*. Note that since we have expanded the cuboid associated with each of the 10 SWB classes, a basin that was successfully assigned to a class by the original SWB method is assigned to the same class when our extended SWB system is used.

The GAGES-II dataset contains a folder called *basinchar_and_report_sept_2011*. Within this folder is a spreadsheet called *gagesII_sept30_2011_var_desc.xlsx* that describes the 354 variables available in GAGES-II. For the 9067 basins in CONUS, the values of these variables are in the file *gagesII_sept30_2011_conterm.xlsx*. For the 255 basins in Alaska, Hawaii, and Puerto Rico, they are in the file: *gagesII_sept30_2011_AKHIPR.xlsx*. The CONUS spreadsheet has a Climate tab with 50 climate variables for each basin, including values for PET (mean annual potential evaporation) and PPTAVG_BASIN (mean annual precipitation). The dimensionless ratio of these is the Budyko aridity index, ϕ. In the variable description spreadsheet, the units of PPTAVG_BASIN and PET are given as *cm* and *mm*, respectively. While the theoretical range for ϕ is $[0, \infty)$, the max attained in the subset of the MOPEX dataset used by Berghuijs et al. (2014) was 5.3 while the max attained in the GAGES-II CONUS dataset was 5.71. This max occurs for the site named “Amargosa River at Tecopa, CA", in the Mojave Desert near Death Valley (site ID = 10251300). The aridity index only exceeds a value of 4 for 8 GAGES-II basins.

While the GAGES-II CONUS dataset does not directly provide the “seasonality timing index”, δ_p, we were able to compute it using available information as follows. Woods (2009), cited by Berghuijs et al. (2014), introduced a “dimensionless mean temperature” denoted as $\overline{T}^*$ in his equation (5). It is defined as:

$$\overline{T}^* = \frac{(\overline{T} - T_0)}{|\Delta T|} \tag{1}$$

Here, $\overline{T}$ is the annual average temperature, T_0 is the rain-to-snow temperature threshold (taken to be 1 degree C), and $|\Delta T|$ is the amplitude of a sine curve fitted to the monthly average temperature for a year. Plots were made to confirm that a sine curve

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

indeed provides a good fit to monthly average temperatures. Equation (13) in Woods (2009) is given by

$$f_s\left(\overline{T}^*, \delta_p^*\right) = \frac{1}{2} - \frac{\arcsin\left(\overline{T}^*\right)}{\pi} - \frac{\delta_p^*}{\pi}\sqrt{1-\left(\overline{T}^*\right)^2}. \tag{2}$$

and expresses the snowiness index, f_s, as a function of both $\overline{T}^*$ and the seasonality index, δ_p. Solving for δ_p we get

$$\delta_p^*\left(f_s, \overline{T}^*\right) = \frac{\pi\left(1/2 - f_s\right) - \arcsin\left(\overline{T}^*\right)}{\sqrt{1-\left(\overline{T}^*\right)^2}}. \tag{3}$$

The GAGES-II CONUS dataset provides the snowiness index, f_s, as SNOW_PCT_PRECIP, in units of percent. It also provides a 30-year average air temperature for each month (from 800 m PRISM data) with variable names like JAN_TMP7100_DEGC. These were used to compute $\overline{T}$ and then $|\Delta T|$ from the average min and max monthly temperature for each basin. These, in turn, were used to compute $\overline{T}^*$ and finally δ_p. For most basins, July was the warmest month and January was coldest. But for 11 basins either December or February was the coldest month, and for 237 basins either June or August was the warmest. (As a side note, GAGES-II CONUS also provides variables called T_MAX_BASIN and T_MIN_BASIN, but their difference was not the same as $|\Delta T|$ computed from monthly values.) It turns out that the cases where $\overline{T}^* = -1$ or $\overline{T}^* = 1$ require special attention because then the δ_p term drops out in (2) and we can't solve for it. For those cases, first notice that equation (2) implies that $f_s = 0$ when $\overline{T}^* = 1$ (since arcsin(1) $= \pi/2$) and $f_s = 1$ when $\overline{T}* = -1$ (since arcsin(-1) $= -\pi/2$). Then observe that the limit of $\delta_p(0, \overline{T}^*)$ as $\overline{T}^*$ goes to 1 is 1 and the limit of $\delta_p(1, \overline{T}^*)$ as $\overline{T}^*$ goes to -1 is -1. These limits can be computed using L'Hopital's rule or with Mathematica.

An important typo was found in equation (6a) in the paper by Berghuijs et al. (2014). That equation is supposed to match equation (13) in Woods (2009) whom they cite (our equation 2), but there is a missing power of 2 on $\overline{T}^*$ inside the square root which is important. A plot of the snowiness index vs. $\overline{T}^*$ (equation 2) over the range $[-1, 1]$ makes more sense for the correct equation, as seen in Figure 9.

Figure 10 shows the result of applying this extended SWB classification scheme to the 9067 basins in the GAGES-II CONUS dataset. One thing that stands out is the distribution of red dots associated with the class A1. According to the descriptions in Table 3 (Berghuijs et al., 2014), one feature of these basins is that the vegetation should be mainly coniferous. The red dots in southern Missouri match with the known distribution of (mostly shortleaf pine) conifers in Missouri. The belt of red dots in Kentucky roughly match a region known as The Knobs, with oak-pine forests. Similarly, the red dots just west of the mountain ranges of the Pacific Northwest also match with coniferous forests. The single red dot in central Nevada appears to coincide with the Humboldt-Toiyabe National Forest, the largest national forest in the lower 48 states, also dominated by conifers. The single red dot in north central Utah appears to coincide with a group of coniferous national forests, like the Uinta-Wasatch-Cache National Forest. The red dots in the Idaho Panhandle match the conifer-dominated Idaho Panhandle National Forest.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

There is also a line of red dots (mostly in western North Carolina and Virginia) that closely match the distribution of Table Mountain pine trees in the Appalachians.

For the GAGES-II CONUS dataset, the number of basins in each extended class is: A1:532, A2:227, A3:25, B1:998, B2:110, C1:1327, C2:472, D1:442, D2:3266, and D3:1668. Since the number of basins in the D2 and D3 classes is so large compared to other classes, it may be worthwhile to extend the SWB method further by adding 2 additional classes, say E1 and E2. Classes E1 and E2 would have the same upper and lower bounds for aridity and snow fraction as D2 and D3. However, instead of extending the upper bound on the seasonality index for classes D2 and D3 to 1, these classes would be added to the right of classes D2 and D3, as seen in the lower part of Figure 6. Whereas PET and precipitation would be weakly in phase for classes D2 and D3, they would be more strongly in phase for classes E1 and E2.

This extended SWB classification is defined in the Python utility: *swb_utils.py*, and its application to the GAGES-II CONUS dataset is given in the utility: *gages2_utils.py*. In *gages2_utils.py*, the functions related to SWB classification are: *compute_swb_classes()*, *get_snow_precip_fraction()*, *get_precip_timing_index()*, *get_aridity_index()*, and *get_swb_points()*.

7 Requirements for a River Basin Classification System

This effort is motivated by the goal of creating a classification system that will allow model agnostic development efforts such as NextGen to reliably match basins with appropriate or "best available" NextGen formulations/models. A number of general characteristics are recommended, leveraging the resources assembled in this effort.

(1) A basin classification system should have a manageable number of basin classes for a given classification objective. Recall that the HLR system has 20 classes, the Seasonal Water Balance (SWB) system has 10 classes, and GAGES-II makes use of and provides 12 ecoregion classes.

(2) A basin classification system should be based on the fundamental concept of *hydrologic similarity*, which must consider which combination of hydrologic processes as well as climate forcings are controlling the hydrologic response of a given basin.

(3) A basin collection for model testing should contain multiple gauged basins from each "basin class", of various sizes, ideally offering sufficient sample sizes for statistical robustness of analysis in each class.

(4) Ideally, it should be possible to assign a basin to a "basin class" even if it is ungauged, as is possible with HLR. Methods based on "hydrologic signatures", metrics, or indices require an observed or estimated time series of variables such as discharge.

(5) There are many basins that are included in multiple collections, such as CAMELS, MOPEX, NOAA RFC basins, and USGS NWIS basins. It is desirable to select basins that span several of the basin collections listed in the section: Overview

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

of Existing River Basin Data Collections. For example, Reynolds Creek is both a USDA-ARS experimental basin and a CZO basin.

(6) Ideally, a classification system should allow us to reliably pair basins with appropriate or "best available" NextGen formulations/models. Depending on the specific objectives of the modeling, several different classification systems may be derived, weighting attributes differently and screening basins for specific characteristics. For instance, for catchment modeling to support flash flood prediction, the collection may exclude basins which lack sub-daily flow measurements; or to focus on modeling in pristine catchments, the collection may exclude basins with significant impairment.

8 Python Utilities for Working with River Basin Data Collections

As explained in the Introduction, a set of Python utilities was created for the purpose of collating, checking, cleaning, augmenting, and extracting information from the various river basin data collections that comprise the HARBOR repository. These Python utilities currently reside in the **utils/ngen** folder within the TopoFlow 3.6 repo (Peckham, 2024c) (since they make use of some TopoFlow utilities) but they may be moved into their own GitHub repo soon. TopoFlow 3.6 is a spatial, hydrologic model consisting of many BMI-enabled process components (Peckham et al., 2017; Peckham, 2024b). BMI stands for Basic Model Interface, a model interface standard introduced by Peckham et al. (2013).

This **ngen** folder contains many Python source code files with names like: **camels_utils.py**, **mopex_utils.py**, **gages2_utils.py**, **usgs_utils.py**, and **rfc_utils.py**. These are Python modules that contain functions for working with the CAMELS, MOPEX, GAGES-II, USGS NWIS, and NOAA RFC datasets, respectively. These utilities were used to prepare augmented TSV (tab-separated value) files for each of the datasets. Each module is interspersed with detailed comments and is cleanly written. They make use of functions in the files: **data_utils.py** (for general data processing tasks) and **shape_utils.py** (for scanning ESRI shapefiles to extract information such as the geographic bounding box (i.e., minlat, maxlat, minlon, maxlon). After preparing an augmented TSV file for each dataset with these utilities, the **collate()** function in the file **collate_basins.py** was used to create a single TSV file with selected attributes from all of the datasets. This TSV file is in the folder called **__Collated** in this repo.

Each Python module contains functions that are helpful for working with a particular dataset. For example, USGS site names almost always contain abbreviations, which are often difficult to decipher. The module *usgs_utils.py* contains an innovative function called *expand_usgs_site_name()* that expands virtually every abbreviation used in these names. This expanded, or "long name" is included in the master TSV file.

The basic algorithm used to collate data from the various data collections is to first create an augmented TSV file with key information for each data collection. These TSV files are in a folder called **_New** within each basin collection folder in the repo. Each such TSV file includes a column with the USGS site ID (available in most cases), and each TSV file is sorted on

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

values in that column. All of the USGS site IDs found in all of these TSV files are combined into a master list of IDs that is also sorted. The *collate()* function in *collate_basins.py* next opens all of these TSV files for reading and then steps through each ID in the master ID list. If the current ID matches the next ID in any of the TSV files, information from those files is read and merged to create a more detailed record for the site with that ID, and this record is then written to a "master TSV file" called *collated_basins_all.tsv* in the repo's **__Collated** folder. This is reminiscent of how the well-known "merge sort" algorithm works.

Note that the algorithm described in the previous paragraph implicitly assumes that a given USGS site ID only occurs once in any of the TSV files. However, it was discovered that in the TSV file created from the NOAA HADS-to-USGS crosswalk, there are 35 USGS Site IDs that occur twice and 2 that occur 3 times (03612600, 04228500). For the 35 that occur twice, all information is identical for 2 of them, while the rest are mapped to different NWS Location IDs and GOES IDs, but have identical name, latitude, and longitude. However, on NOAA websites of the form: *https://water.noaa.gov/gauges/BCBW1*, the names, latitudes, and longitudes for these tend to be different. It appears that in such cases the additional NWS location IDs refer to **auxiliary gages** that are near the **base gage**. For example, for the USGS site ID: 12396500, the USGS site name is: *PEND OREILLE RIVER BELOW BOX CANYON NEAR IONE, WA*, the 2 NWS location IDs are: BCAW1 and BCBW1, and their NOAA names are: *Pend Oreille River below Box Canyon Dam* and *Pend Oreille River below Box Canyon Dam Auxiliary Station*. For USGS site ID: 03612600, the USGS site name is: *OHIO RIVER AT OLMSTED, IL*, the 2 NWS location IDs are: OLMI2 and OLTI2, and their NOAA names are: *Ohio River at Olmsted Lock and Dam Headwater* and *Ohio River at Olmsted Lock and Dam Tailwater*.

In order to provide as much information as possible for each site and to avoid missing data, multiple sources of information — obtained by multiple methods such as web-scraping, APIs, and alternate data sets — were incorporated in a pre-processing phase. In this phase, Python dictionaries were created to map a key (like a site ID) to a record (another dictionary) containing multiple fields of data associated with that key. Python dictionaries use a very efficient hashing algorithm and therefore allow rapid retrieval of stored data for a site within the *collate()* function. In some cases (e.g., when pulling data from an API, or scraping web pages), the creation of these Python dictionaries was relatively slow, so they were saved for rapid import in Python PKL files in the **_New** folder associated with a given data collection. With regard to issues like missing or non-conformant data, obtaining complete data for NOAA RFC sites proved to be the most challenging.

In addition to these Python utilities, QGIS (QGIS, 2024) an open-source GIS application, was used to view ESRI shapefile attribute tables and to save them to CSV format. It was also used to investigate oddities in the datasets and to apply a point-in-polygon algorithm to determine HUC12 codes for USGS site IDs. When QGIS was used, a README file was added to the repo with a step-by-step workflow description for the sake of reproducibility.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

9 Conclusions

The extensive effort described in this paper has gathered and systematized river gaging datasets that were or are collected by many agencies with different descriptive schema and made available via dissimilar web services. To the extent possible, all the variability of these sources has been preserved in a final master listing of the collected basin metadata, along with further descriptive tags. The result is one of the major outcomes of the work, a collated basin metadata file (in TSV format), *collated_basins_all.tsv*, found in the **__Collated** folder of the HARBOR repo. It contains 51 attributes for 30,717 river basins from a variety of different river basin data collections. It can be read into any spreadsheet program and searched, or information can be sorted on the data in any column to isolate the data needed to answer a particular question. Python programs can easily be written (e.g., the included module *subset.py*) that search the TSV file for particular combinations of attributes. A good starting point for a diverse collection of river basins with numerous attributes, classifications, and shapefiles, etc. is the GAGES-II SB3 dataset, which consists of 1947 basins. For each of these, HARBOR provides an HLR code and an extended SWB classification, and many also have a hydrograph-based classification based on NWM3 analysis. Keep in mind that all the basins in the SB3 subset of GAGES-II are among the so-called "Reference" basins, which means they have relatively low human impact.

Another key goal of this project was to make water data collected by US federal agencies more accessible. Looking at these various river basin data collections all together rather than separately, and in a historical context, leads to a deeper appreciation of their value in addressing many different societal problems.

Code and data availability. These Python utilities currently reside in a folder within the TopoFlow 3.6 repo, at: Peckham (2024c) and make use of some of the TopoFlow utilities. TopoFlow 3.6 is a spatial, hydrologic model consisting of many BMI-enabled process components. See Peckham et al. (2017). The data sets themselves, and a TSV file with information gleaned from all of them, were collected into a new GitHub repository that is currently available at Peckham (2024a).

Appendix A: Terminology

The purpose of this section is to define some terminology, some of which is agency-specific, that is helpful for understanding the information provided by the various datasets.

A1 USGS Terminology

The words "site" and "station" are used interchangeably to refer to a particular "point-type" location where a measurement or prediction is made, such as a streamgage, well, or meteorological station. See Dupré et al. (2013). The USGS assigns a unique number or "site ID" to each site that consists of at least 8 and up to 15 digits. The first two digits are called the "part number" and refer to a particular large basin/region in the USA. These range from 01 for North Atlantic Slope basins up 16 for basins in Hawaii. The next 6 digits are called the "downstream-order number" and attempt to indicate relative position in the

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

downstream direction within the region defined by the part number. Gaps between site IDs are reserved for future use, but in areas of high station density, more than 8 digits may be used. See the About Sites section of the USGS-NWIS (2024a). Each site is also assigned a name, but these names are heavily abbreviated. (Code is included in the Python utility *usgs_utils.py* that expands most of these abbreviations, and these expanded site names are also given in the "collated" TSV file in the HARBOR repo.). In addition to an ID and a name, each site also has a "site type". Site types are strings such as: Stream, Atmosphere, Well, Spring, and Lake. In order to determine what was measured at a particular site, you can look at the NWIS *data_types_cd* (data types code). This is a 30-character array, where each character is one of: A (for Active data collection site), I (for inactive or discontinued data collection site), O (for inventory site only), or N (possibly for "not or never collected at this site"). Each position in this string corresponds to a different type of measurement, but only the first 16 positions are currently used. For example, the first character in the array corresponds to: *Stage or water-levels – continuous*; see USGS-NWIS (2024c) for more information. It appears that the USGS establishes a stage-discharge *rating curve* for each site that collects stage data, so if the first character is an A, then presumably an estimate of discharge is available. **(CHECK)** It is possible that an agency other than the USGS may have primary responsibility for a site and then a "site agency" may be specified. Each site also has a Hydrologic Unit Code, or HUC. A HUC has at least 2 digits, and each additional digit (up to 16 total) indicates another level of spatial detail in a hierarchy of watershed boundary polygons for the US. So, a site with a given HUC number is contained within a US watershed polygon that has that number. For example, the average area of a HUC12 watershed polygon is 15 to 62 square km. The HUC system only applies to the US, although other national and international watershed boundary datasets exist. See Seaber et al. (1987) and USGS-HUC (2024a, b).

If you know the USGS Site ID for a site, then you can obtain information about and historical and real-time data for that site by entering a URL of the following form in your browser:

http://waterdata.usgs.gov/nwis/nwisman/?site_no=01010000

And to obtain real-time streamflow (or discharge) information for a USGS Site ID, you can use a URL of the form:

https://waterwatch.usgs.gov/index.php?mt=real&st=01010000

Similarly, you can get information about all the sites that lie within a HUC8 watershed polygon by entering a URL of the following form in your browser:

https://water.usgs.gov/lookup/getwatershed?10250001

Unfortunately, you can currently only obtain information this way for 8-digit HUC numbers. Both the site ID and HUC URLs are included in the collated basin TSV file of the NextGen repo. The USGS makes its data available online through NWIS, the National Water Information System. NWIS will be described in more detail in a subsequent section. When you download data from the NWIS website, it is in a text file format called the *RDB format* (for relational database), which consists of a header with a standardized format to describe the downloaded data followed by a sequence of newline-separated data records with tab-separated fields. USGS data can also be downloaded from the Water Quality Portal, which provides USGS NWIS data, USDA STEWARDS data, and EPA STORETS data. Long before NWIS, STEWARDS, or STORET existed, starting in 1971, the USGS used a system called WATSTORE (National Water Data Storage and Retrieval System). See Hutchinson (1975) for

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

its documentation, which can still be helpful and contains parameter code listings. It is included in the HARBOR repo in the folder *USGS_NWIS_Web/Docs/WATSTORE*.

A2 NOAA-NWS Terminology

In NOAA NWS terminology, the word "location" may be used as a synonym for site. However, NWS has its own system of assigning unique IDs to a site/location. The "NWS location ID" (or NWSLID or "Handbook 5 ID" or SID or 5-char ID) typically has a base length of 5 alphanumeric characters. (Three or more additional characters are often added by the RFCs according to their own needs, but this practice is not standardized.) These IDs start with a 3-letter prefix, which is usually an abbreviation of a city or station name, followed by a two-character alphanumeric SID state code. The SID state code starts with the first letter of the state's name, followed by a single digit from 1 to 9. See NOAA (2024db, Appendix A) for SID state codes, e.g., Alaska = A2, Connecticut = C3, Florida = F1, Michigan = M4. NOAA or USGS sites that transmit their data to a GOES satellite (called DCPs or Data Collection Platforms) also have a unique GOES (or NESDIS) ID. The NWS has the entire US divided into 13 large regions/basins, and there is a River Forecast Center (RFC) for each of these regions. RFC names are abbreviated to 5 letters, where the last three letters are RFC and the first 2 are an abbreviation like MB for Missouri Basin. An RFC region contains many County Warning Areas (CWAs), which each contain a Weather Forecast Office (WFO). However, CWAs may intersect more than one RFC region near the RFC boundaries. Each CWA (and its associated WFO) has a unique 3-letter abbreviation, such as ABQ for Albuquerque, NM, or SLC for Salt Lake City, UT. A CWA can be further divided into Hydrologic Service Areas (HSAs), which also have 3-letter abbreviations. In principle, each NWS location ID should "fall within" or "belong" to a particular RFC, CWA, and HSA, but a complete list of these mappings is not readily available on the Web. Just as the USGS provides a URL for each USGS Site ID, NOAA's NWPS (National Water Prediction Service) provides a URL with information for each NWS location ID of the following form:

https://water.noaa.gov/gauges/rbum7

where *rbum7* can be any NWS location ID. The NWPS also provides a URL with information for all sites within a given WFO of the form:

https://water.noaa.gov/area/EWX

where EWX can be replaced with any 3-letter WFO ID. In addition, each RFC has its own website with an interactive map where information can be viewed for individual locations. The URLs all have the form: *https://water.noaa.gov/rfc/wgrfc*. The NWPS also just released a new API (NOAA, 2024e) which can be used to obtain information associated with a particular USGS Site ID or NWS location ID.

Note that HADS (Hydrometeorological Automated Data System) is a real-time data acquisition and data distribution system operated by the NWS Office of Dissemination. There is a NOAA HADS-to-USGS crosswalk table (NOAA-NCEP, 2024) that provides mappings for some NWS Location IDs to equivalent USGS Site IDs and GOES (or NESDIS) IDs.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

While NWS does not have a "site type" equivalent, sometimes a 5-character PEDTS code is provided and information about the quantity that is observed at the site is embedded in this code. Parts of the code correspond to the letters in PEDTS: PE=Physical Element, D=Duration Code, T=Type Code, and S=Source Code. PEDTS codes are defined in the "SHEF manual" (NOAA, 2012). SHEF stands for Standard Hydrometeorological Exchange Format and was introduced by the NWS for interagency data sharing.

A3 USDA Terminology

The USDA (US Department of Agriculture) is the parent department for many other agencies, including the Agricultural Research Service (ARS), the Forest Service (USFS), the Natural Resource Conservation Service (NRCS), and the Farm Service Agency (FSA). Due to the importance of watersheds to both agriculture, forests, and other rural land, the USDA also engages in numerous hydrologic monitoring efforts, with a strong focus on variables other than streamflow. These include the ARS Experimental Watershed Network, the USFS Watershed Condition Framework, and the NRCS Snow Survey and Water Supply Forecasting (SSWSF) program, which runs the SCAN (Soil Climate Analysis Network) and SNOTEL (Snow Telemetry) projects. The ARS Experimental Watershed effort is now part of CEAP (Conservation Effects Assessment Project) and its Watershed Assessment Studies Network (WASN) led by NRCS (CEAP, 2021). Until recently, it was possible to browse and download data for the Experimental Watersheds in a given US state by entering a URL of the following form into a browser:

https://hrsl.ba.ars.usda.gov/wdc/md.htm|

where "md" is replaced by the 2-letter abbreviation for that state. Many of these pages have been discontinued, but nonetheless contained useful links to all related data for a given site and can still be viewed with the "Internet Archive Wayback Machine". It is possible that alternative access methods have emerged, but the current status is unclear. Much of the data resource appears to be available from a USDA website called "Ag Data Commons". However, the main way to obtain USDA data is by using a system called STEWARDS (USDA-STEWARDS, 2024). STEWARDS does for the USDA-ARS what NWIS does for the USGS. Similarly, WQX (previously STORET) plays this same role for the USEPA data. NWIS, STEWARDS, and WQX data are also available through the Water Quality Portal (WQP), a service sponsored by the USGS and EPA. See WQP (2024).

A4 Latitudes, Longitudes, Elevations, Areas, and Datums

While the concepts of latitude, longitude, and elevation may seem straightforward, when it comes to determining these values for a given location on the surface of the Earth, one realizes that it isn't really that simple. Even the concepts of latitude and longitude depend on whether one is treating the Earth as a sphere or an ellipsoid, and on deciding on where the center of the Earth is located. A horizontal datum provides the mechanism for determining latitude and longitude at a point, and a vertical datum does the same for elevation. One way to create a horizontal datum (which was used for NAD 1927, where NAD stands for North American Datum) is to pick an origin (e.g., Meades Ranch in Kansas) and to decide on the latitude and longitude of that origin and then to work outward with careful surveys to create a network of control points that have latitudes and longitudes consistent with (or relative to) that origin. A survey (or geodetic) marker is often anchored in rock at a datum control point. So, when reporting the latitude and longitude for a point-type location, such as a streamgage, it is important to know the corresponding

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

datum. In the US, NAD 1927, NAD 1983, and WGS 1984 are all used. The position of a given point can differ by as much as 500 feet depending on which datum is used. WGS 1984 is a modern datum associated with GPS (Global Positioning System), which uses satellite-based technology. Similarly, elevations must be specified relative to a particular vertical datum, and the issues are very similar. For the US, the main ones are the National Geodetic Vertical Datum of 1929, and the North American Vertical Datum of 1988. NOAA has plans to release new National Geodetic Survey datums in 2025. NOAA also provides online conversion tools for both horizontal and vertical datums: NOAA-NCAT (2024) and NOAA-VDATUM (2024). Note that it is typically not necessary to specify more than 5 digits after the decimal for latitudes and longitudes, because 5-digits provide accuracy to within about 1 meter. See Wikipedia (2024a, b, c). Keep in mind also that a reported latitude and longitude may be for a streamgage (watershed outlet), a watershed centroid, a study region centroid, or the center of a bounding box.

In many cases, a river basin data provider will provide an ESRI shapefile for each watershed polygon in a given collection. This vector data file contains the latitudes and longitudes of a sequence of ordered points on the watershed boundary. For the acquisition of associated topographic, soil, or climate forcing data it is often helpful to know the "geographic bounding box" for a basin. This is given by just 4 numbers, namely the min and max or both latitude and longitude for all of the points on the watershed boundary (plus some padding). This bounding box has been provided whenever possible in the collated basin TSV file of the HARBOR Repo.

Appendix B: A Brief History of the USGS and NOAA

Following a recommendation by the National Academy of Sciences, the USGS was established by Congress on March 3, 1879, with Clarence King as its first Director. John Wesley Powell was the second Director of the USGS, from 1881-1894. Under his leadership, the first USGS stream gages were installed on the Rio Grande River (near Embudo, New Mexico) as part of the National Streamgaging Program in 1889. The fist automatic water-stage recorders at streamgages were installed as early as 1912. The rich history of the USGS from its inception in 1879 up to 1989 has been well-documented by Rabbitt (1989). In addition, the history of the USGS Water Resources Division (WRD) has also been documented in a set of 8 volumes, although only the last 4 volumes have been published (Ferguson, 1990; Hudson and Cragwall, 1996; Biesecker et al., 2000; Blakey et al., 2005).

While NOAA was not formed until 1970, it brought several much older agencies together under one administration, including the Survey of the Coast (created in 1807 while Thomas Jefferson was president), the Weather Bureau (from 1870), and the U.S. Commission of Fish and Fisheries (from 1871). Also in 1970, the Weather Bureau was renamed to the National Weather Service (NWS). See NOAA (2024c). In 1975, GOES-1, the first geostationary satellite owned and operated by NOAA was launched. By 1982, NOAA's GOES DCS (data collection system) had numerous users, including federal agencies, with the Department of the Interior (e.g., USGS), the Department of Commerce (e.g., NOAA), and the Department of Defense (e.g., USACE) being the three biggest users (about 85% of total). These developments led to the current state of affairs where data

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

is (1) automatically collected throughout a network of stream gauges, (2) transmitted via a directional Yagi-Uda antenna to a geostationary NOAA GOES satellite, (3) relayed to intermediate servers at Wallops Island, VA, and finally (4) made available to users on USGS and NOAA websites. While most stream gages are operated and maintained by the USGS, they are typically funded through a partnership with other federal, state, local and tribal entities. An abbreviated timeline of NOAA's history is given in Table A1.

The USGS and NOAA have a long history of working together and both have responsibility for managing and modeling the water resources of the US (along with the USACE, EPA, USDA, and others). Both agencies also have a history of working together with academic researchers, passing knowledge back and forth between "research" and "operations". NOAA's 16 Cooperative Institutes (CIs) provide a key example of this collaboration, with CIROH (Cooperative Institute for Research to Operations in Hydrology) being the most recent CI and the first to focus on hydrology. Both the USGS and NOAA's OWP have adopted the Basic Model Interface (BMI) standard for model interoperability which originated in the academic community under NSF funding (Peckham et al., 2013). Both agencies are also working together on the "Hydrofabric" that underpins the Next Generation Water Resources Modeling Framework (NextGen) effort; see Blodgett et al. (2023). NextGen will be used for National Water Model versions after 3.2.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Year	Important Event in NOAA's History
1807	Survey of the Coast is created under President Thomas Jefferson.
1870	Weather Bureau is created by Congress under President Ulysses S. Grant.
1871	U.S. Commission of Fish and Fisheries is created.
1940	Weather Bureau is transferred from USDA to the Department of Commerce (DoC) by President Franklin D. Roosevelt, due to its growing importance to the aviation industry.
1940	The Office of Hydrology (OH) is created within the Weather Bureau.
1946	The first two River Forecast Centers (RFC) are established by the Weather Bureau, one in Cincinnati, OH and one in Kansas City, MO. Several RFCs were opened, merged, and closed between 1946 and 1979, resulting in the 13 RFCs we have today.
1951	Weather Bureau initiates aviation radio, which leads to NOAA Weather Radio in 1970. It is expanded under Vice President Al Gore to cover 95% of the US population.
1967	The first Cooperative Institute, the Cooperative Institute for Research in Environmental Science (CIRES) is created at the University of Colorado Boulder.
1970	NOAA is officially formed and brings several other agencies under one administration. The Weather Bureau is renamed to the National Weather Service (NWS).
1975	NOAA's first geostationary satellite, GOES-1, is launched.
1982	The GOES DSC has many users, including DoC/NOAA, DoI/USGS, DoD/USACE.
1992	The OH Hydrology Laboratory (HL) begins distributed modeling research.
1997	Office of Hydrology begins the Advanced Hydrologic Prediction Service (AHPS) program.
2000	Office of Hydrology (OH) is renamed to Office of Hydrologic Development (OHD).
2008	OHD begins the Community Hydrologic Prediction System (CHPS) based on Delft-FEWS.
2014	NOAA's National Water Center (NWC) opens in Tuscaloosa, AL in mid-2014.
2015	The NWC has its ribbon-cutting ceremony in mid-2015.
2016	NOAA's Office of Water Prediction (OWP) is created from OHD.
2016	NOAA's National Water Model (NWM) version 1 is released in August.
2019	NWM version 2.0 is released, followed by version 2.1 in 2021, and version 3.0 in 2023.
2021	Work begins on Next Generation Water Resources Modeling Framework, to become NWM 4.0
2022	NOAA's 16th cooperative institute is created, CIROH (Cooperative Institute for Research to Operations in Hydrology), the first CI to focus on hydrology.
2024	NOAA's NWS launches a new website called the National Water Prediction Service (NWPS) that merges data from AHPS and OWP. This includes a new API for data access.

Table A1: **Abbreviated NOAA History Timeline.** A list of events in the history of NOAA that help to put the current work into context. A much more detailed and interactive timeline for the National Weather Service is given at: NOAA (2024da).

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Author contributions. Peckham developed the HARBOR repository, wrote the associated Python utilities and the paper itself. Jennings supervised the project as lead hydrologist and reviewed and edited the paper. Wu reviewed the HARBOR repository and provided hydrograph classifications for many basins from the National Water Model version 3. Wood helped to resolve lingering issues with the NOAA River Forecast Center data sets. Bolotin provided the sections on hydrologic signatures and data-driven machine-learning methods. All authors reviewed the paper for accuracy and readability and provided edits and references.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to thank Missouri Basin RFC employees Kevin Low, Eugene Derner, Jessica Brooks, Scott Dummer, Gregg Schalk, and Maren Stoflet for helpful discussions and for providing a mapping of runoff zones to their corresponding basins with ESRI shapefiles. We would also like to thank (1) Daryl E. Herzmann at Iowa State University for providing assistance with the Iowa Environmental Mesonet (IEM) website which provided some missing data for some of NOAA's RFCs, and (2) Matej Durcik of SAHRA (University of Arizona) for providing access to the MOPEX data set.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

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, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Addor, N., Nearing, G., Prieto, C., Newman, A. J., LeVine, N., and Clark, M. P.: A ranking of hydrological signatures based on their predictability in space, Water Resources Research, 54, 8792–8812, https://doi.org/10.1029/2018WR022606, 2018.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5946, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., and Poulin, A.: HYSETS — A 14425 watershed hydrometeorological sandbox over North America, https://osf.io/rpc3w/, accessed: 2024-08-05, 2011.

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M., Ameli, A., and Poulin, A.: A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds, Scientific Data, 7, 243, https://doi.org/10.1038/s41597-020-00583-2, 2020.

Berghuijs, W. R., Sivapalan, M., Woods, R. A., and Savenije, H. H. G.: Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales, Water Resources Research, 50, 5638–5661, https://doi.org/10.1002/2014WR015692, 2014.

Biesecker, J. E., Blakey, J. F., Feltz, H. R., and George, J. R.: A history of the Water Resources Division, U.S. Geological Survey, Volume VII, 1966-79 — Integrating the disciplines, U.S. Government Printing Office, https://doi.org/10.3133/70039522, 345 pp., 2000.

Blackburn-Lynch, W., Agouridis, C. T., and Barton, C. D.: Development of regional curves for Hydrologic Landscape Regions (HLR) in the contiguous United States, Journal of the American Water Resources Association, 53, 903–928, https://doi.org/10.1111/1752-1688.12540, 2017.

Blakey, J. F., Biesecker, J. E., Feltz, H. R., Kantrowitz, I. H., and Yong, L. E.: A history of the Water Resources Division, U.S. Geological Survey: Volume VIII, 1979-94, U.S. Government Printing Office, https://doi.org/10.3133/70047300, 599 pp., 2005.

Blodgett, D., Johnson, J. M., and Bock, A.: Generating a reference flow network with improved connectivity to support durable data integration and reproducibility in the conterminous US, Environmental Modelling and Software, 165, 105 726, https://doi.org/10.1016/j.envsoft.2023.105726, 2023.

Bolotin, L. A. and McMillan, H. K.: A hydrologic signature approach to analysing wildfire impacts on overland flow, Hydrological Processes, (in press), 2024.

Bolotin, L. A., Haces-Garcia, F., Liao, M., and Liu, Q.: Automated decision support for model selection in the NextGen National Water Model, in: CUAHSI Technical Report 18, pp. 7–15, CUAHSI, in coordination with the NOAA OWP Summer Institute program, 2022.

Brantley, S. L., McDowell, W. H., Dietrich, W., White, T. S., Kumar, P., Anderson, S. P., Chorover, J., Lohse, K. A., Bales, R. C., Richter, D. D., Grant, G., and Gaillardet, J.: Designing a network of critical zone observatories to explore the living skin of the terrestrial Earth, Earth Surface Dynamics, 5, 841–860, https://doi.org/10.5194/esurf-5-841-2017, 2017.

Brooks, P. D., Troch, P. A., Durcik, M., Gallo, E., and Schlegel, M.: Quantifying regional scale ecosystem response to changes in precipitation: Not all rain is created equal, Water Resources Research, 47, W00J08, https://doi.org/10.1029/2010WR009762, 13 pp., 2011.

Budyko, M. I.: The Heat Balance of the Earth's Surface, US Department of Commerce, Weather Bureau, Washington, DC, https://doi.org/10.1080/00385417.1961.10770761, 259 pp., translated by Nina A. Stepanova, 1958.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Cook, T.: Big benefits from experimental watersheds, https://eos.org/research-spotlights/big-benefits-from-experimental-watersheds, EOS, Feb. 18, 2021; accessed: 2024-08-05, 2021.

Coopersmith, E., Yaeger, M. A., Ye, S., Cheng, L., and Sivapalan, M.: Exploring the physical controls of regional patterns of flow duration curves — Part 3: A catchment classification system based on regime curve indicators, Hydrology and Earth System Sciences, 16, 4467–4482, https://doi.org/10.5194/hess-16-4467-2012, 2012.

Cosgrove, B., Gochis, D., Flowers, T., Dugger, A., Ogden, F., Graziano, T., Clark, E., Cabell, R., Casiday, N., Cui, Z., Eicher, K., Fall, G., Feng, X., Fitzgerald, K., Frazier, N., George, C., Gibbs, R., Hernandez, L., Johnson, D., Jones, R., Karsten, L., Kefelegn, H., Kitzmiller, D., Lee, H., Liu, Y., Mashriqui, H., Mattern, D., McCluskey, A., McCreight, J. L., McDaniel, R., Midekisa, A., Newman, A., Pan, L., Pham, C., RafieeiNasab, A., Rasmussen, R., Read, L., Rezaeianzadeh, M., Salas, F., Sang, D., Sampson, K., Schneider, T., Shi, Q., Sood, G., Wood, A., Wu, W., Yates, D., Yu, W., and Zhang, Y.: NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling, Journal of the American Water Resources Association, 60, 247–272, https://doi.org/10.1111/1752-1688.13184, 2024.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320, 3–17, https://doi.org/10.1016/j.jhydrol.2005.07.031, 2006.

Dupré, D. H., Scott, J. C., Clark, M. L., Canova, M. G., and Stoker, Y.: User's manual for the National Water Information System of the U.S. Geological Survey: Water-Quality System, Version 5.0, Open-File Report 2013-1054, USGS, US Dept. of the Interior, https://pubs.usgs.gov/of/2013/1054/pdf/OFR2013-1054_NWIS_ver5.pdf, 771 pp., 2013.

Durcik, M., Troch, P., Brooks, P. D., and Sivapalan, M.: Data for Hydrologic Synthesis Project 2009, University of Arizona, https://arizona.app.box.com/s/ygm97gumftyneskn9gbggqx9sq3othnx, accessed: 2024-08-14; original URL was nero.hwr.arizona.edu/mopex/, 2009.

Eberts, S. M., Woodside, M. D., Landers, M. N., and Wagner, C. R.: Monitoring the pulse of our nation's rivers and streams — The US Geological Survey Streamgaging Network, https://doi.org/10.3133/fs20183081, accessed: 2024-08-06, 2018.

Falcone, J. A.: GAGES–II: Geospatial attributes of Gages for Evaluating Streamflow: U.S. Geological Survey data release, https://doi.org/10.5066/P96CPHOT, accessed: 2024-08-05, 2011.

Falcone, J. A.: U.S. Geological Survey GAGES-II time series data from consistent sources of land use, water use, agriculture, timber activities, dam removals, and other historical anthropogenic influences: U.S. Geological Survey data release, https://doi.org/10.5066/F7HQ3XS4, accessed: 2024-08-12, 2017.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, Ecology, 91, 621, https://doi.org/10.1890/09-0889.1, Data Paper, Ecological Archives E091-045, 2010.

Ferguson, G. E.: A history of the Water Resources Division, U.S. Geological Survey, Volume V, July 1, 1947, to April 30, 1957, U.S. Government Printing Office, https://doi.org/10.3133/7000089, 309 pp., 1990.

Franklin, J. F., Bledsoe, C. S., and Callahan, J. T.: Contributions of the Long-Term Ecological Research Program, BioScience, 40, 509–523, https://doi.org/10.2307/1311319, 1990.

Gnann, S. J., Howden, N. J. K., and Woods, R. A.: Hydrological signatures describing the translation of climate seasonality into streamflow seasonality, Hydrology and Earth System Sciences, 24, 561–580, https://doi.org/10.5194/hess-24-561-2020, 2020.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Goodrich, D. C., Starks, P. J., Schnabel, R. R., et al.: Effective use of USDA-ARS experimental watersheds, in: Agricultural Research Service Conference on Hydrology, edited by Richardson, C. W., Rango, A., Owens, L. B., and Lane, L. J., Publication 1994-5, pp. 35–46, U.S. Department of Agriculture., Agriculture Research Service, Denver, Colorado, 1994.

Goodrich, D. C., Marks, D., Seyfried, M. S., Keefer, T. O., Unkrich, C. L., Anson, E. A., Clark, P. E., Flerchinger, G. N., Hamerlynck, E. P., Hardegree, S. P., Heilman, P., Holifield-Collins, C., Moran, M. S., Nearing, M. A., Nichols, M. H., Pierson, F. B., Scott, R. L., Stone, J. J., Vactor, S. S. V., Winstral, A. H., and Wong, J. K.: Utilizing long-term ARS data to compare and contrast hydroclimatic trends from snow and rainfall dominated watersheds, in: Proceedings of the 4th Interagency Conference on Research in the Watersheds, pp. 34–39, U.S. Department of Agriculture., Agriculture Research Service, Fairbanks, Alaska, https://www.tucson.ars.ag.gov/unit/Publications/PDFfiles/2142.pdf, 2011.

Goodrich, D. C., Heilman, P., Moran, S., Garbrecht, J., Marks, D., Bosch, D., Steiner, J., Sadler, J., Romkens, M., Harmel, D., Kleinman, P., Gunter, S., and Walbridge, M.: The USDA-ARS Experimental Watershed Network - Evolution, lessons learned, and moving forward, in: Headwaters to estuaries: advances in watershed science and management — Proceedings of the Fifth Interagency Conference on Research in the Watersheds, edited by Stringer, C. E., Krauss, K. W., and Latimer, J. S., pp. 54–60, U.S. Department of Agriculture, Forest Service, Southern Research Station, North Charleston, SC, https://research.fs.usda.gov/treesearch/50873, 2016.

Goodrich, D. C., Heilman, P., Anderson, M., Baffaut, C., Bonta, J., Bosch, D., Bryant, R., Cosh, M., Endale, D., Veith, T. L., Havens, S. C., Hedrick, A., Kleinman, P. J., Langendoen, E. J., McCarty, G., Moorman, T., Marks, D., Pierson, F., Rigby, J. R., Schomberg, H., Starks, P., Steiner, J., Strickland, T., and Tsegaye, T.: The USDA-ARS Experimental Watershed Network: Evolution, lessons learned, societal benefits, and moving forward, Water Resources Research, 57, https://doi.org/10.1029/2019WR026473, 27 pp., 2020.

Gosz, J. R., Waide, R. B., and Magnuson, J. J.: Twenty-Eight Years of the US-LTER Program: Experience, Results, and Research Questions, in: Long-Term Ecological Research: Between Theory and Application, edited by Müller, F., Baessler, C., Schubert, H., and Klotz, S., pp. 59–74, Springer, Dordrecht, https://doi.org/10.1007/978-90-481-8782-9_5, 2010.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

Hobbie, J. E., Carpenter, S. R., Grimm, N. B., Gosz, J. R., and Seastedt, T. R.: The US Long Term Ecological Research Program, BioScience, 53, 21–32, https://doi.org/10.1641/0006-3568(2003)053[0021:TULTER]2.0.CO;2, 2008.

Hudson, H. H. and Cragwall, Jr., J. S.: A history of the Water Resources Division, U.S. Geological Survey: Volume VI, May 1, 1957, to June 30, 1966 - The Years of Change, U.S. Government Printing Office, https://doi.org/10.3133/7000084, 559 pp., 1996.

Hutchinson, N. E.: WATSTORE: National Water Data Storage and Retrieval System of the U.S. Geological Survey, User's Guide, Open-File Report 75-426, U.S. Government Printing Office, https://doi.org/10.3133/ofr75426, 956 pp., 1975.

Kennard, M. J., Pusey, B. J., Olden, J. D., MacKay, S. J., Stein, J. L., and Marsh, N.: Classification of natural flow regimes in Australia to support environmental flow management, Freshwater Biology, 55, 171–193, https://doi.org/10.1111/j.1365-2427.2009.02307.x, 2010.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Nature Scientific Data, 10, 1081–1100, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Mai, J., Craig, J. R., Tolson, B. A., and Arsenault, R.: The sensitivity of simulated streamflow to individual hydrologic processes across North America, Nature Communications, 13, https://doi.org/10.1038/s41467-022-28010-7, 2022.

McMillan, H. K.: A review of hydrologic signatures and their applications, WIREs Water, 8, https://doi.org/10.1002/wat2.1499, 23 pp., 2020a.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

McMillan, H. K.: Linking hydrologic signatures to hydrologic processes: A review, Hydrological Processes, 34, 1393–1409, https://doi.org/10.1002/hyp.13632, 2020b.

Müller, F., Baessler, C., Schubert, H., and Klotz, S., eds.: Long Term Ecological Research: Between Theory and Application, Springer, https://doi.org/10.1007/978-90-481-8782-9, 450 pp., 2010.

NCAR: CAMELS (Catchment Attributes and Meteorology for Large-sample Studies), Data Downloads, https://ral.ucar.edu/solutions/products/camels, accessed: 2024-08-05, 2024.

NEON: History, https://www.neonscience.org/about/overview/history, accessed: 2024-08-05, 2024a.

NEON: Explore field sites, https://www.neonscience.org/field-sites/explore-field-sites, accessed: 2024-08-05, 2024b.

NEON: NEON Website, https://www.neonscience.org/, accessed: 2024-08-05, 2024c.

NEON: NEON R Package, https://www.neonscience.org/resources/learning-hub/tutorials/neondatastackr, accessed: 2024-08-05, 2024d.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., and Arnold, J. R.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

NOAA: Standard Hydrometeorological Exchange Format (SHEF) Manual, National Weather Service Manual 10-944, July 5, 150 pp., https://www.weather.gov/media/mdl/SHEF_CodeManual_5July2012.pdf, accessed: 2024-08-06, 2012.

NOAA: NOAA River Forecast Center Website, https://water.weather.gov/ahps/rfc/rfc.php, accessed: 2024-08-05, 2024b.

NOAA: NOAA Heritage Website, https://www.noaa.gov/heritage, accessed: 2024-08-05, 2024c.

NOAA: NOAA History Timeline, https://vlab.noaa.gov/web/nws-heritage/explore-nws-history, accessed: 2024-08-05, 2024da.

NOAA: Site Identifiers: National Weather Service Instruction 30-1204, NWS Directive System (NDS) 30-12: Maintenance, Logistics and Facilities (30), Configuration Management (12), 27 pp., https://www.weather.gov/media/directives/030_pdfs/pd03012004curr.pdf, updated: August 5, 2024, accessed: 2024-08-06, 2024db.

NOAA: National Weather Prediction Service API, NOAA, Office of Water Prediction, https://api.water.noaa.gov/nwps/v1/docs/, accessed: 2024-08-09, 2024e.

NOAA-NCAT: NGS Coordinate Conversion and Transformation Tool (NCAT), NOAA National Geodetic Survey (NGS), https://geodesy.noaa.gov/NCAT/, accessed: 2024-08-09, 2024.

NOAA-NCEP: USGS-NWSLI Cross Reference, All Sites, https://hads.ncep.noaa.gov/USGS/, accessed: 2024-08-09, 2024.

NOAA-VDATUM: Vertical Datum Transformation, NOAA National Geodetic Survey (NGS) and Office of Coast Survey (OCS) and Center for Operational Oceanographic Products and Services (CO-OPS) , https://vdatum.noaa.gov/, accessed: 2024-08-09, 2024.

Normand, A. E.: U.S. Geological Survey (USGS) Streamgaging Network: Overview and issues for Congress, Congressional Research Service R45695, 32 pp., https://sgp.fas.org/crs/misc/R45695.pdf, accessed: 2024-08-06, 2021.

NSF-CZO: CZO site characteristics 2018, https://czo-archive.criticalzone.org/images/national/associated-files/1National/CZO_Attributes_Table.pdf, accessed: 2024-08-05, 2021.

Ogden, F., Avant, B., Bartel, R., Blodgett, D., Clark, E., Coon, E., Cosgrove, B., Cui, S., Kindl da Cunha, L., Farthing, M., Flowers, T., Frame, J., Frazier, N., Graziano, T., Gutenson, J., Johnson, D., McDaniel, R., Moulton, J., Loney, D., Peckham, S., Mattern, D., Jennings, K., Williamson, M., Savant, G., Tubbs, C., Garrett, J., Wood, A., and Johnson, J.: The Next Generation Water Resources Modeling Framework: Open Source, Standards Based, Community Accessible, Model Interoperability for Large Scale Water Prediction, in: AGU Fall Meeting Abstracts, vol. 2021, pp. H43D–01, 2021.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Over, T. M., Farmer, W. H., and Russell, A. M.: Refinement of a regression-based method for prediction of flow-duration curves of daily streamflow in the conterminous United States: USGS Science Investigations Report 2018-5072, U.S. Geological Survey, Reston, VA, https://doi.org/10.3133/sir20185072, 34 pp., 2018.

Peckham, S. D.: HARBOR GitHub Repo: Harmonized Attributes for River Basins in One Repo, https://github.com/peckhams/nextgen_basin_repo, accessed: 2024-08-05, 2024a.

Peckham, S. D.: TopoFlow version 3.6 GitHub Repository, https://github.com/peckhams/topoflow36, accessed: 2024-08-06, 2024b.

Peckham, S. D.: TopoFlow version 3.6 GitHub Repository, NextGen utilities, https://github.com/peckhams/topoflow36/tree/master/topoflow/utils/ngen, accessed: 2024-08-06, 2024c.

Peckham, S. D., Hutton, E. W. H., and Norris, B.: A component-based approach to integrated modeling in the geosciences: The Design of CSDMS, Computers & Geosciences, 53, 3–12, https://doi.org/10.1016/j.cageo.2012.04.002, special issue: Modeling for Environmental Change, 2013.

Peckham, S. D., Stoica, M., Jafarov, E. E., Endalamaw, A., and Bolton, W. R.: Reproducible, component-based modeling with TopoFlow, a spatial hydrologic modeling toolkit, Earth and Space Science, 4, 377–394, https://doi.org/10.1002/2016EA000237, special issue: Geoscience Papers of the Future, 2017.

QGIS: QGIS Website, A free and open source Geographic Information System (GIS), https://qgis.org, accessed: 2024-08-09, 2024.

Rabbitt, M. C.: The United States Geological Survey: 1879-1989, USGS Circular 1050, U.S. Geological Survey, https://doi.org/10.3133/cir1050, report, 52 pp., 1989.

Russell, A. M., Over, T. M., and Farmer, W. H.: Streamflow, flow-duration curves, basin characteristics, and regression models of flow-duration curves for selected streamgages in the conterminous United States: U.S. Geological Survey data release, https://doi.org/10.5066/F70G3JC4, accessed: 2024-08-05; see Child Item 3: Selected basin characteristics for all GAGES-II CONUS reference gages, 2018.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrology and Earth Science Sciences, 15, 2895–2911, https://doi.org/10.5194/hess-15-2895-2011, 2011.

Schaake, J., Cong, S., and Duan, Q.: The US MOPEX Data Set, in: Large sample basin experiments for hydrological model parameterization: Results of the model parameter experiment — MOPEX, pp. 9–28, International Association of Hydrological Sciences, IAHS Publication 307, ISSN 0144-7815, 2006.

Seaber, P. R., Kapinos, F. P., and Knapp, G. L.: Hydrologic Unit Maps, U.S. Geological Survey Water Supply Paper 2294, 66 pp., https://pubs.usgs.gov/wsp/wsp2294/, accessed: 2024-08-05, 1987.

Slack, J. R. and Landwehr, J. M.: Hydro-Climatic Data Network (HCDN): A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874-1988; U.S. Geological Survey, Open-file Report 92-129, 200 pp., Reston, VA , https://doi.org/10.3133/ofr92129, accessed: 2024-08-05, 1992a.

Slack, J. R. and Landwehr, J. M.: Contents of the Hydro-Climatic Data Network (HCDN), https://pubs.usgs.gov/of/1992/ofr92-129/conthcdn.html, accessed: 2024-08-06, 1992b.

Slack, J. R., Lumb, A. M., and Landwehr, J. M.: Hydro-Climatic Data Network (HCDN) – A USGS streamflow data set for the U.S. for the study of climate fluctuations: U.S. Geological Survey data release, https://doi.org/10.5066/P9HP0WFJ, accessed: 2024-08-06, 1994.

Troch, P., Dwivedi, R., Neto, A. A. M., Liu, T., Roy, T., Valdés-Pineda, R., Durcik, M., Arciniega-Esparza, S., and Breña-Naranjo, J. A.: Data for the catchment-scale groundwater recharge and vegetation water use efficiency estimation, HydroShare, http://www.hydroshare.org/resource/99d5c1a238134ea6b8b767a65f440cb7, 2018.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

UNWI: Cross-CZO – Streamflow / Discharge — USGS and USDA Data Resources – National – (1985-2017), HydroShare, http://www.hydroshare.org/resource/a60faac4c73649efb4736de7ec587f7c, Upper Northwest Interceptor (UNWI) System, accessed: 2024-08-05, 2019.

USDA, A.: USDA Agricultural Research Service Water Database, https://data.nal.usda.gov/dataset/ars-water-database, accessed: 2024-08-06, 2024.

USDA-NWCS: Natural Resources Conservation Service, Air and Water Database Public Reports, NRCS National Water Climate Data Center, Active SCAN Stations as of 2024-August-06, https://wcc.sc.egov.usda.gov/nwcc/yearcount?network=scan&counttype=statelist, accessed: 2024-08-06, 2024a.

USDA-NWCS: USDA Natural Resources Conservation Service, Soil Climate Analysis Network (SCAN), https://www.nrcs.usda.gov/resources/data-and-reports/soil-climate-analysis-network, accessed: 2024-08-06, 2024b.

USDA-NWCS: Natural Resources Conservation Service, Air and Water Database Public Reports, NRCS National Water Climate Data Center, Active SNOTEL Stations as of 2024-August-06, https://wcc.sc.egov.usda.gov/nwcc/yearcount?network=sntl&counttype=statelist, accessed: 2024-08-06, 2024c.

USDA-NWCS: USDA Natural Resources Conservation Service, Snow Survey and Water Supply Forecasting System, Snow Telemetry (SNOTEL), https://www.nrcs.usda.gov/programs-initiatives/sswsf-snow-survey-and-water-supply-forecasting-program, accessed: 2024-08-06, 2024d.

USDA-STEWARDS: STEWARDS v4.0: Access to ARC CEAP Benchmark Watershed Data, https://www.nrrig.mwa.ars.usda.gov/stewards/stewards.html, accessed: 2024-08-06, 2024.

USGS: The US Geological Survey Streamgaging Network Website, https://www.usgs.gov/mission-areas/water-resources/science/usgs-streamgaging-network, accessed: 2024-08-06, 2021.

USGS-FPS: Federal Priority Streamgages (FPS) Website, U.S. Geological Survey, https://www.usgs.gov/mission-areas/water-resources/science/federal-priority-streamgages-fps, accessed: 2024-08-06, 2024.

USGS-HUC: Hydrologic Unit Maps, https://water.usgs.gov/GIS/huc.html, accessed: 2024-08-06, 2024a.

USGS-HUC: Hydrologic Unit Codes (HUCs) explained, https://nas.er.usgs.gov/hucs.aspx, accessed: 2024-08-06, 2024b.

USGS-NWIS: USGS NWIS Current Conditions for the Nation, https://waterdata.usgs.gov/nwis/current, accessed: 2024-08-06, 2024a.

USGS-NWIS: USGS NWIS Surface Water Daily Data for the Nation, https://waterdata.usgs.gov/nwis/dv, accessed: 2024-08-06, 2024b.

USGS-NWIS: USGS NWIS Help System: Type of Data Collected, https://help.waterdata.usgs.gov/codes-and-parameters/type-of-data-collected, accessed: 2024-08-14, 2024c.

USGS-NWIS: USGS NWIS Site Inventory for the Nation, https://waterdata.usgs.gov/nwis/inventory, accessed: 2024-08-06, 2024d.

USGS-NWIS: National Water Information System (NWIS): Help System FAQ, https://help.waterdata.usgs.gov/faq, accessed: 2024-08-06, 2024a.

Wikipedia: Decimal degrees, https://en.wikipedia.org/wiki/Decimal_degrees, accessed: 2024-08-05, 2024a.

Wikipedia: Vertical datum, https://en.wikipedia.org/wiki/Vertical_datum, accessed: 2024-08-05, 2024b.

Wikipedia: Horizontal datum, https://en.wikipedia.org/wiki/Geodetic_datum#Horizontal_datum, accessed: 2024-08-05, 2024c.

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al.: The FAIR Guiding Principles for scientific data management and stewardship, Nature Scientific Data, 3, 160 018, https://doi.org/10.1038/sdata.2016.18, 2016.

Winter, T. C.: The concept of hydrologic landscapes, Journal of the American Water Resources Association, 37, 335–349, https://doi.org/10.1111/j.1752-1688.2001.tb00973.x, 2001.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Wlostowski, A. N., Molotch, N., Anderson, S. P., Brantley, S. L., Chorover, J., Dralle, D., Kumar, P., Li, L., Lohse, K. A., Mallard, J. M., McIntosh, J. C., Murphy, S. F., Parrish, E., Safeeq, M., Seyfried, M., Shi, Y., and Harman, C.: Signatures of hydrologic function across the Critical Zone Observatory Network, Water Resources Research, 57, e2019WR026 635, https://doi.org/10.1029/2019WR026635, 2021.

Wolfram Research Inc.: Mathematica, Version 14.1, https://www.wolfram.com/mathematica, Champaign, IL, 2024.

Wolock, D.: Hydrologic landscape regions of the United States: U.S. Geological Survey data release, https://doi.org/10.5066/P9UIFPKW, accessed: 2024-08-05, 2003.

Wolock, D. M., Winter, T. C., and McMahon, G.: Delineation and evaluation of hydrologic-landscape regions in the United States using geographic information system tools and multivariate statistical analyses, Environmental Management, 34, S71–S88, https://doi.org/10.1007/s00267-003-5077-9, 2004.

Woods, R. A.: Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks, Advances in Water Resources, 32, 1465–1481, https://doi.org/10.1016/j.advwatres.2009.06.011, 2009.

WQP: Water Quality Portal (WQP), National Water Quality Monitoring Council, https://www.waterqualitydata.us/, accessed: 2024-08-09, 2024.

Wu, S., Zhao, J., Wang, H., and Sivapalan, M.: Regional patterns and physical controls of streamflow generation across the conterminous United States, Water Resources Research, 57, e2020WR028 086, https://doi.org/10.1029/2020WR028086, 2021.

Ye, S., Yaeger, M., Coopersmith, E., Cheng, L., and Sivapalan, M.: Exploring the physical controls of regional patterns of flow duration curves — Part 2: Role of seasonality, the regime curve, and associated process controls, Hydrology and Earth System Sciences, 16, 4447–4465, https://doi.org/10.5194/hess-16-4447-2012, 2012.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

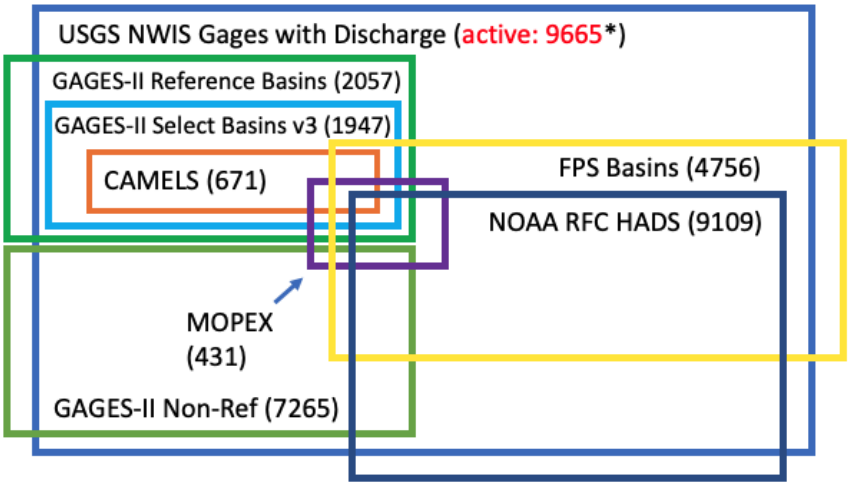

Figure 1. Venn diagram for various river basin data collections.

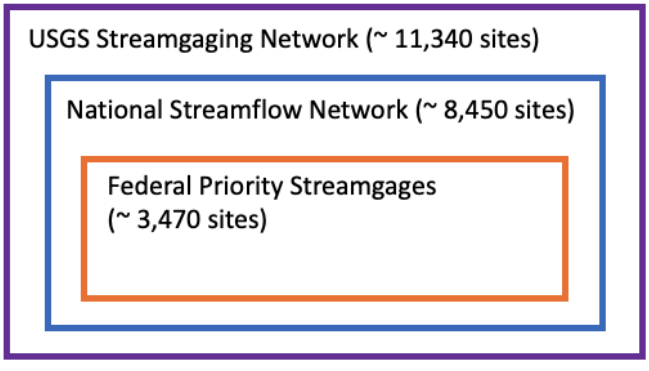

Figure 2. Venn diagram for three USGS river basin data collections, adapted from Normand (2021, Figure 1).

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

Figure 3. The 13 NOAA NWS River Forecast Center (RFC) regions of the USA (from: noaa.gov/jetstream/rfcs).

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

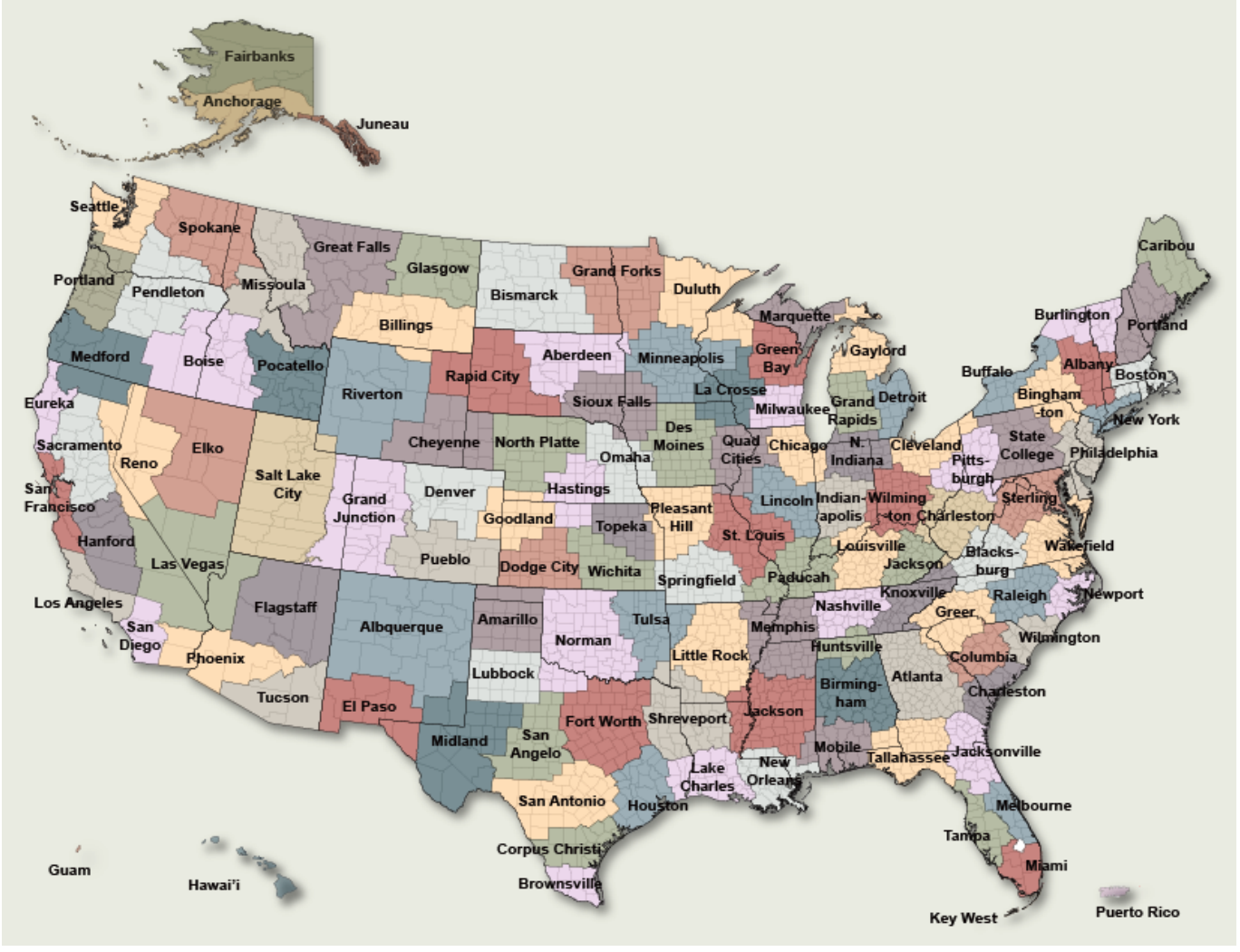

Figure 4. The NOAA NWS Weather Forecast Office (WFO) regions of the USA (from noaa.gov/jetstream/wfos).

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

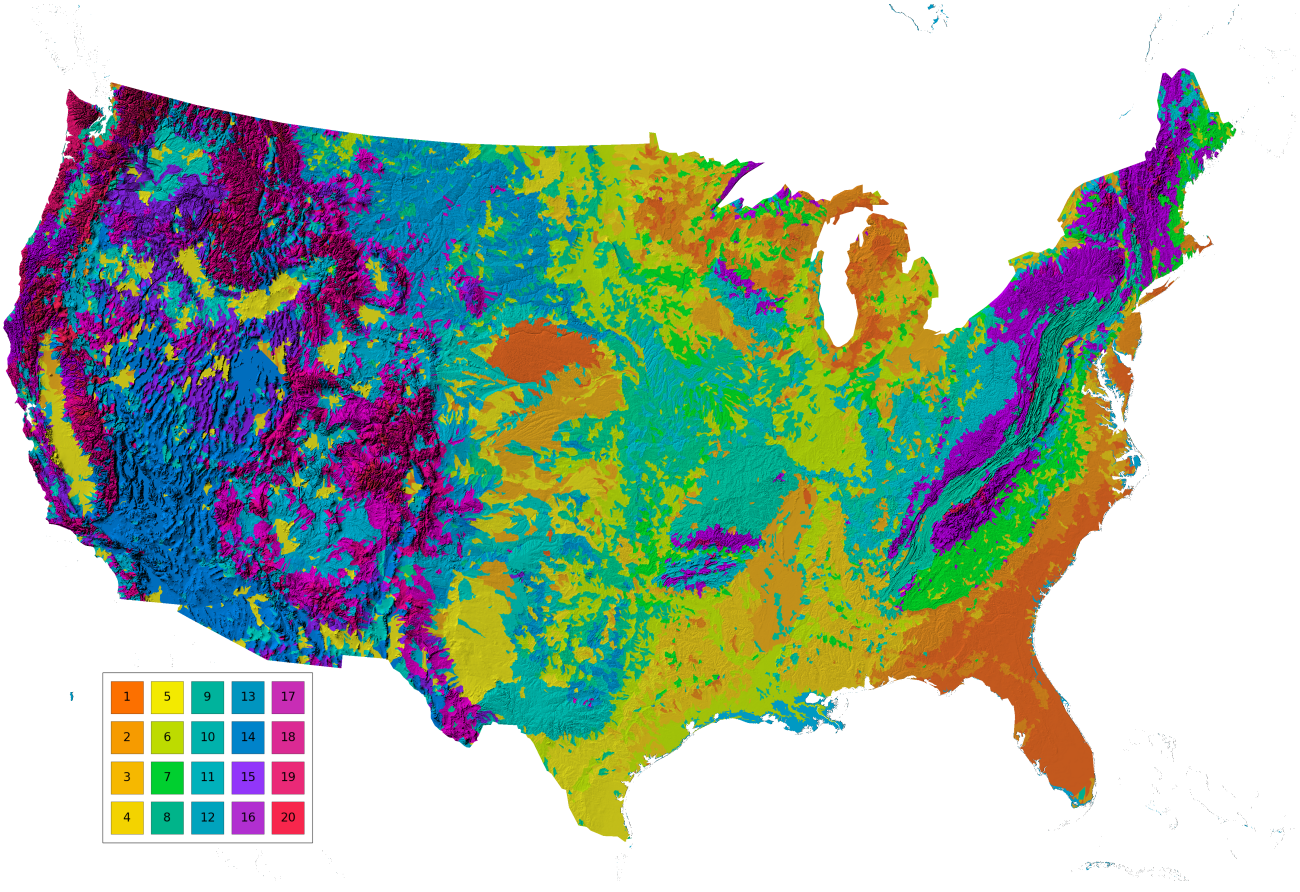

Figure 5. A hillshaded map of the 20 HLR regions within the conterminous US.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

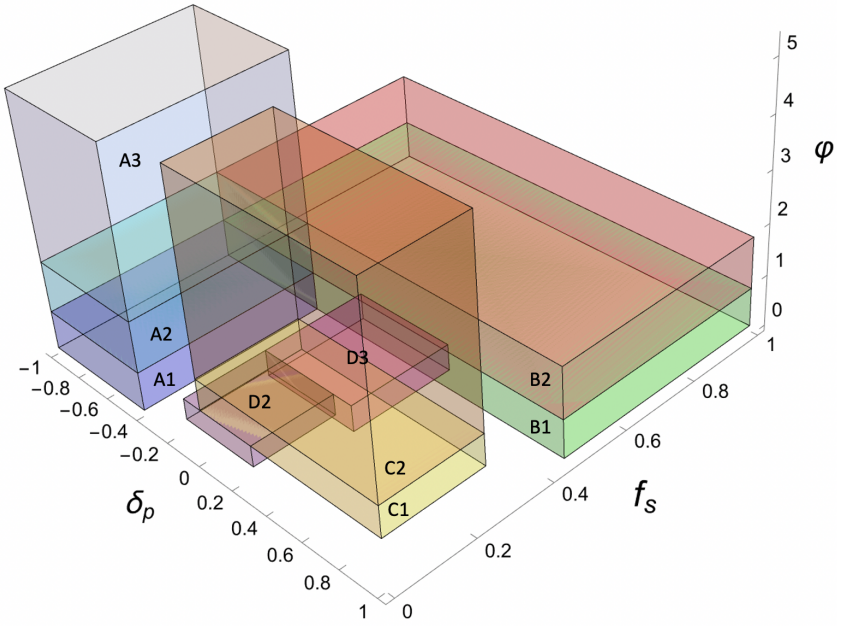

Figure 6. Oblique view of parameter space for the original Seasonal Water Balance method that shows significant gaps.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

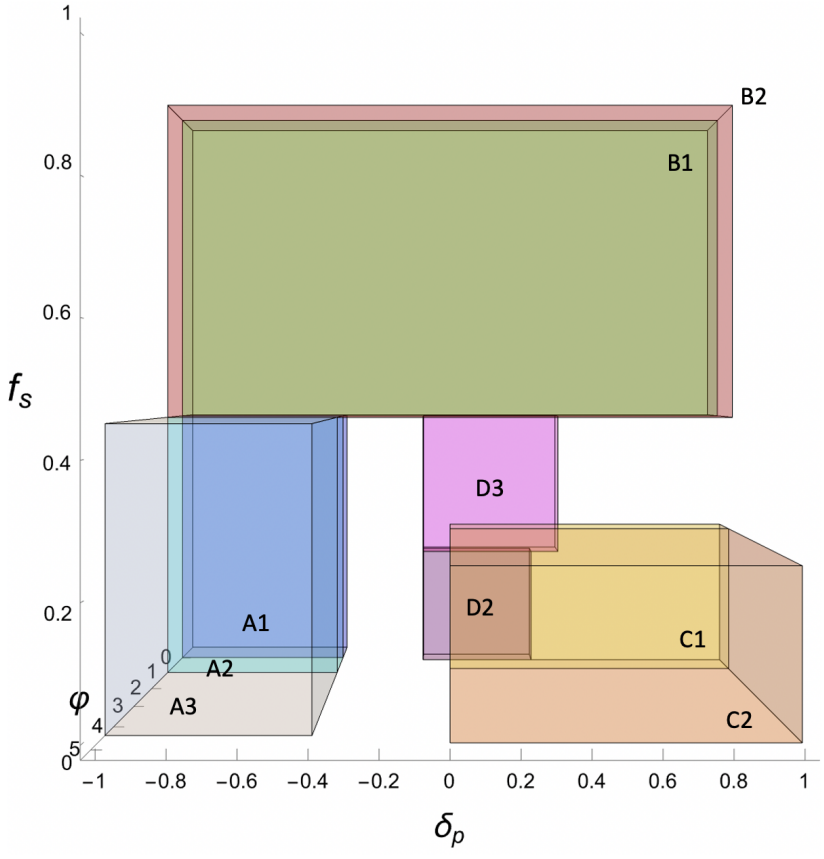

Figure 7. Top view of parameter space for the original Seasonal Water Balance method that shows significant gaps.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

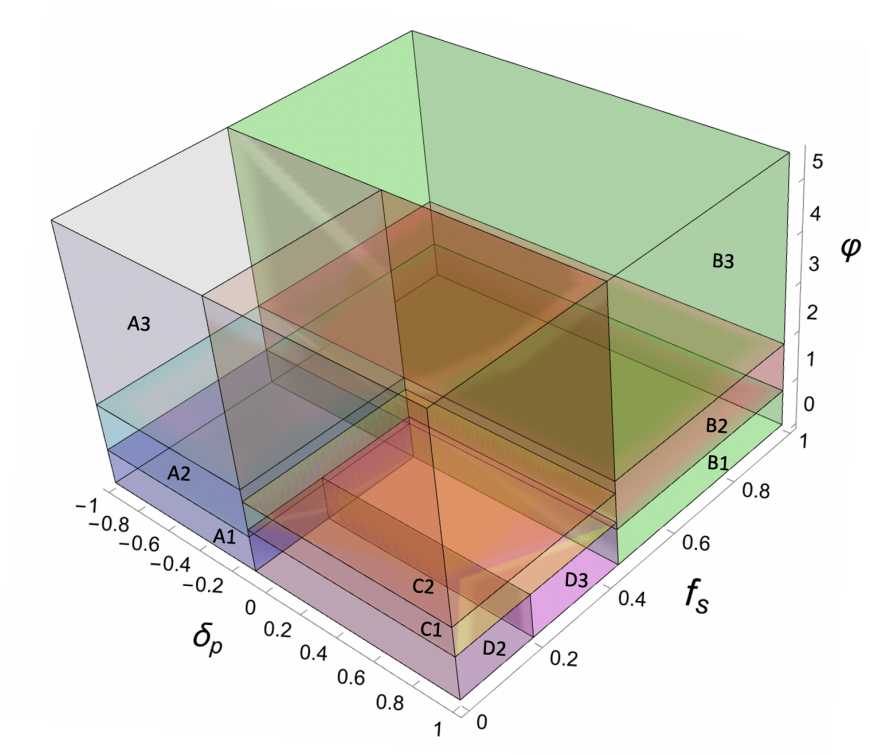

Figure 8. Oblique view of parameter space for the extended Seasonal Water Balance method that has no gaps.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

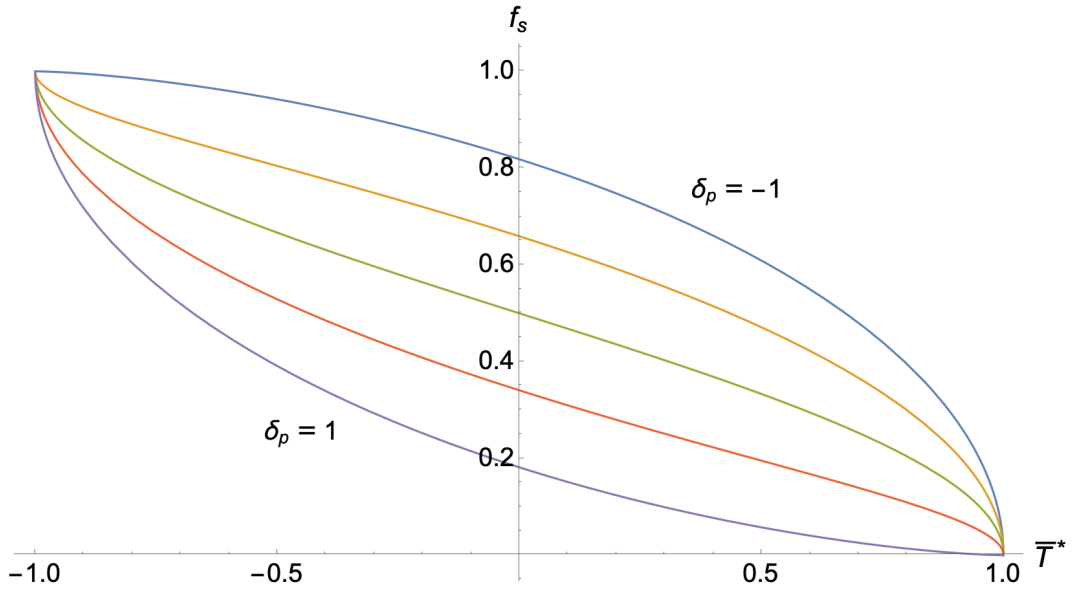

Figure 9. Plot of f_s as a function of $\overline{T}^*$, for δ_p values in $\{-1, -0.5, 0, 0.5, 1\}$. For each curve, f_s decreases from 1 to 0 as $\overline{T}^*$ increases, as expected.

https://doi.org/10.5194/egusphere-2025-5786
Preprint. Discussion started: 7 January 2026

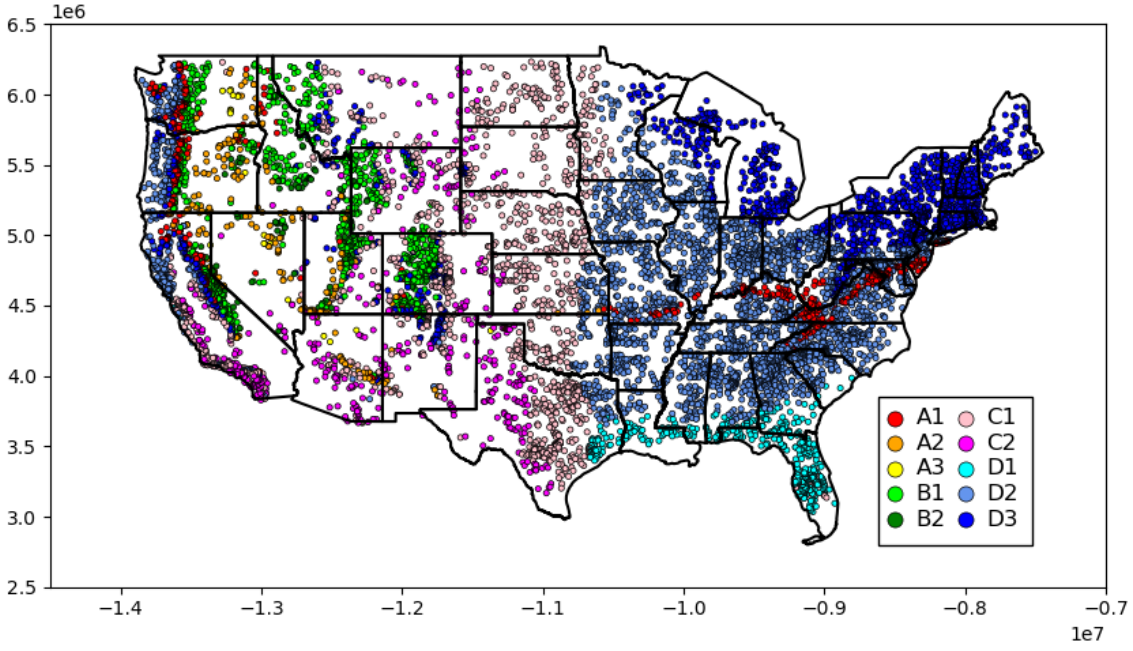

Figure 10. Scatter plot of Extended Seasonal Water Balance (SWB) classifications for GAGES-II CONUS watersheds. The distribution of red dots for class A1 is discussed in the main text. This plot was created with the *create_swb_scatter_plot()* function in *plot_utils.py*.