Catchment Attributes and MEteorology for Large-Sample SPATially distributed modeling studies (CAMELS-SPAT): Streamflow observations, forcing data and geospatial data for hydrologic modeling across North America

Wouter J. M. Knoben[1], Laura Torres Rojas[2], Nathaniel W. Chaney[2], Alain Pietroniro[1], and Martyn P. Clark[1]

[1]ADDRESS
[2]ADDRESS

Abstract. TEXT

1 Introduction

Increases in geospatial data availability and computing power have enabled rapid advances in large-scale (Cloke and Hannah, 2011) and large-sample (Addor et al., 2020) hydrology. A key difference between these fields is the spatial continuity of the study area. Where large-scale studies concern themselves with obtaining predictions across continuous areas, large-sample studies tend to select separate basins within a given area of interest. The large-sample approach strikes a balance between spatial variability and ease of use. This facilitates studies that can be representative of larger spatial regions at a fraction of the computational effort needed to run a large-scale study over the same domain.

A driving force behind the large-sample movement has been the "CAMELS" family of data sets. The original Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) dataset was developed as a two-part initiative. First, basin-averaged meteorological time series were provided for several hundreds of basins across the Contiguous United States (Newman et al., 2015). Second, statistical descriptors (referred to as catchment attributes) of each catchment's hydroclimatic conditions were made available (Addor et al., 2017a). This combined data set has proven useful for various purposes, mainly within the overarching themes of understanding, quantifying and modeling hydroclimatic diversity. The success of this original data set for these purposes has motivated development of multiple (typically national) variants, as well as the aggregated cloud-based CARAVAN collection (Kratzert et al., 2023).

Table 1 provides a brief overview of the main characteristics of various CAMELS(-like) data sets. Because our interest is in hydrologic modeling, we limit this overview to data sets that include meteorologic time series that could serve as input to hydrologic models. A commonality between most of these data sets is a focus on aggregated data: meteorologic forcing data

and catchment attributes are typically provided as basin-averaged values, and the temporal resolution of provided forcing data is almost always at daily time steps. Similarly, most datasets provide a specific selection of forcing variables: precipitation (P) and temperature (T) are always included, as well a potential evapotranspiration (PET) time series or the variables necessary to calculate PET. In modeling terms, these data sets focus strongly on catchment modeling with lumped conceptual models. Such models treat catchments as single (i.e., lumped) entities, are typically run at daily time resolutions, and generally require only time series of P, T and PET to function. Commonly known examples of such models are SAC-SMA (National Weather Service, 2005), HBV (Lindström et al., 1997) and GR4J (Perrin et al., 2003). Such models are computationally cheap but often criticized for their somewhat empirical and spatially lumped nature, and their lack of explicit energy balance calculations.

Spatially-distributed process-based models, such as VIC (Hamman et al., 2018) and SUMMA (Clark et al., 2015a, b), address these concerns but come with the trade-off of increased computational cost and face their own challenges. Notable examples include the definition of appropriate parameter values and questions about the scale-dependency of their constitutive functions. Investigating these models in large-sample studies could provide helpful insights but running such models is not easily possible with most of the data sets listed in Table 1. The clearest exception to this is the LamaH-CE (Klingler et al., 2021) data set, which covers the Upper Danube river basin in Central Europe. LamaH-CE provides data in a semi-distributed, spatially continuous fashion and provides a collection of forcing variables generally associated with process-based modeling approaches. However, the spatially continuous nature of this data set means it is somewhat constrained geographically, covering an area of only 170,000 km^2 (roughly 600 by 300 km), and the Lamah-CE dataset still aggregates data at the sub-basin level. There is a clear gap in the current collection of large-sample hydrologic data sets that (1) enables the use of spatially-distributed process-based models across a wide range of hydroclimatic conditions, and (2) enables studies aimed at investigating spatial heterogeneity at a resolution made possible by the geospatial data sets that underpin the current generation of large-sample hydrology data sets.

In this paper we introduce the CAMELS-SPAT data set ("Catchment Attributes and MEterology for Large-sample Studies for SPATially distributed modeling"). We expand on the original CAMELS data set (Newman et al., 2015; Addor et al., 2017a) in various ways. First, we provide data at native (i.e. gridded), sub-basin and basin levels, instead of treating each catchment only as a lumped entity. Second, we extend the geographical domain of the data set to include Canada, which includes various types of hydrologically challenging landscapes not included in the original CAMELS data set (e.g., glaciated basins, regions with extensive permafrost, arctic deserts). Third, we provide a wider range of forcing variables at a temporal resolution (i.e., hourly) suitable for process-based modeling. Compared to LamaH-CE, our main contributions can be found in the wider range of hydroclimatic conditions found across the United States and Canada, and the inclusion of forcing and geospatial data at their native (non-aggregated) resolution.

Table 1. Overview of large-sample data sets aimed at hydrologic modeling. Data sets are listed chronologically.

Data set	Coverage			Resolution		Forcing data	
	Temporal	Spatial Region	# basins	Temporal	Spatial	# products	Variables
MOPEX[1,2]	1948-2003	Contiguous US	438	Daily	Basin-averaged	Station observations within basins	Precipitation, climatic potential evaporation, maximum air temperature, minimum air temperature
CANOPEX[1]	Varies per basin	Canada	698	Daily	Basin-averaged	2 (1 station, 1 gridded)	Precipitation, maximum temperature, minimum temperature
CAMELS[1]	Varies per forcing dataset	Contiguous US	671	Daily	Basin-averaged; per elevation band	3	Precipitation, maximum temperature, minimum temperature, shortwave downward radiation, day length, vapor pressure
CAMELS-CL[1]	Varies per basin	Chile	516	Daily	Basin-averaged	Multiple, depending on variable	Precipitation, maximum temperature, mean temperature, minimum temperature, potential evapotranspiration
HYSETS[1]	Varies per basin	North America	14425	Daily	Basin-averaged	7	Precipitation, maximum air temperature, minimum air temperature
CAMELS-BR[1]	1981-2018	Brazil	897	Daily	Basin-averaged	Multiple, depending on variable	Precipitation, maximum temperature, average temperature, minimum temperature, potential evapotranspiration, actual evapotranspiration
CAMELS-GB[1]	1970-2015	Great Britain	671	Daily	Basin-averaged	1	Precipitation, average temperature, potential evapotranspiration, potential evapotranspiration with interception correction, wind speed, specific humidity, downward shortwave radiation, longwave radiation
CABra[1]	1980-2010	Brazil	785	Daily	Basin-averaged	3	Precipitation, maximum temperature, minimum temperature, solar radiation, 2m wind speed, potential evapotranspiration (3 estimates), actual evapotranspiration
CAMELS-AUS[1]	Varies per forcing dataset and variable	Australia	222	Daily	Basin-averaged	Multiple, depending on variable	Precipitation, maximum temperature, minimum temperature, potential evapotranspiration (4 estimates), actual evapotranspiration, solar radiation, vapor pressure, vapor pressure deficit, relative humidity at time of maximum temperature, relative humidity at time of minimum temperature, mean sea level pressure
LamaH-CE[1,3]	1981-2019	Central Europe	859	Daily, hourly	Basin-averaged at three basin levels	1	Precipitation, 2m air temperature, 10m wind in U-direction, 10m wind in V-direction, net solar radiation at the surface, net thermal radiation at the surface, surface pressure, total evapotranspiration
CCAM[1]	1990-2020	China	4911	Daily	Basin-averaged	1	Precipitation, 2m mean temperature, ground surface temperature, potential evapotranspiration, measured evaporation, ground pressure, relative humidity, 2m wind speed, sunshine duration,
CAMELS-CH[1,4]	1981-2020	Switzerland and surrounding areas	331	Daily	Basin-averaged	1 (used to derive extra variables)	Precipitation, maximum temperature, mean temperature, minimum temperature, relative sunshine duration

[1] References: MOPEX (Schaake et al., 2006), CANOPEX (Arsenault et al., 2016), CAMELS (Newman et al., 2015; Addor et al., 2017a) , CAMELS-CL (Alvarez-Garreton et al., 2018), HYSETS (Arsenault et al., 2020), CAMELS-BR (Chagas et al., 2020), CAMELS-GB (Coxon et al., 2020), CABra (Almagro et al., 2021), CAMELS-AUS (Fowler et al., 2021), LamaH-CE (Klingler et al., 2021), CCAM (Hao et al., 2021), CAMELS-CH (Höge et al., 2023).

[2] MOPEX forcing variables as currently available on https://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/Us_438_Daily/.

[3] LamaH-CE basins are spatially connected.

[4] CAMELS-CH forcing variables derived from the core forcing include: precipitation, mean temperature, global radiation, sunshine duration, wind speed, relative humidity, potential evapotranspiration, actual evapotranspiration, intercepted evapotranspiration

2 Design considerations and outcomes

Our goal with this data set is to enable studies that investigate spatial heterogeneity across a wide variety of catchments, with a specific focus on spatially-distributed process-based modeling. We also envision this data set to be used to compare the performance of these models to their more empirical counterparts. As a result, we need to process a variety of data sources at various levels. We provide further detail about these requirements in the following sub-sections, as needed. Our general methodology for creating CAMELS-SPAT is as follows:

1. Define an initial set of basins of potential interest, covering the United States and Canada;
2. Create consistent basin delineations for all basins identified under (1);
3. Obtain and process streamflow observations for the basins identified under (1), removing those basins for which no streamflow data can be found;
4. Obtain and process meteorological forcing data for the basins identified under (3);
5. Obtain and process geospatial data sets (e.g. data describing each basins climate, vegetation, land use, topography, soil and geology) for the basins identified under (3);
6. Remove a number of very large basins from the basins identified under (3), and divide the remaining basins into various sub-datasets, based on disk space considerations.

Figure 1 shows a visual summary of the main steps and decision points in this process, and each step is explained in more detail in the following subsections. For the reader's benefit, we present a brief description of our methods as well as the results for each of these steps together in the same subsection, instead of splitting these out into dedicated Methods and Outcomes sections. Code necessary to reproduce our methods is freely available (see "Code and Data Availability" statement, Section 5). This code resource could also be used to obtain data for the basins that were removed in Step 6 listed above.

2.1 Basin preselection

2.1.1 Context

We impose two initial constraints on the basins we will consider including in this data set. First, we have chosen to focus this dataset on (near-)natural basins. Human impacts on the earth system are critically important but substantially complicate hydrologic behaviour and are typically difficult to quantify. Such impacts include, but are not limited to: (i) the construction of water management structures such as dams and drainage ditches at the local level, of which the location and size are difficult to ascertain and usually unreported in the continental scale data sets CAMELS-SPAT relies on; (ii) the construction of large water management infrastructure such as diversions and reservoirs, which may appear in continental scale data sets but for which operating procedures are typically unknown; (iii) surface and groundwater abstractions for e.g. agricultural and industrial use, for which initial abstraction and return volumes are typically unknown. Second, we require the availability of at least some

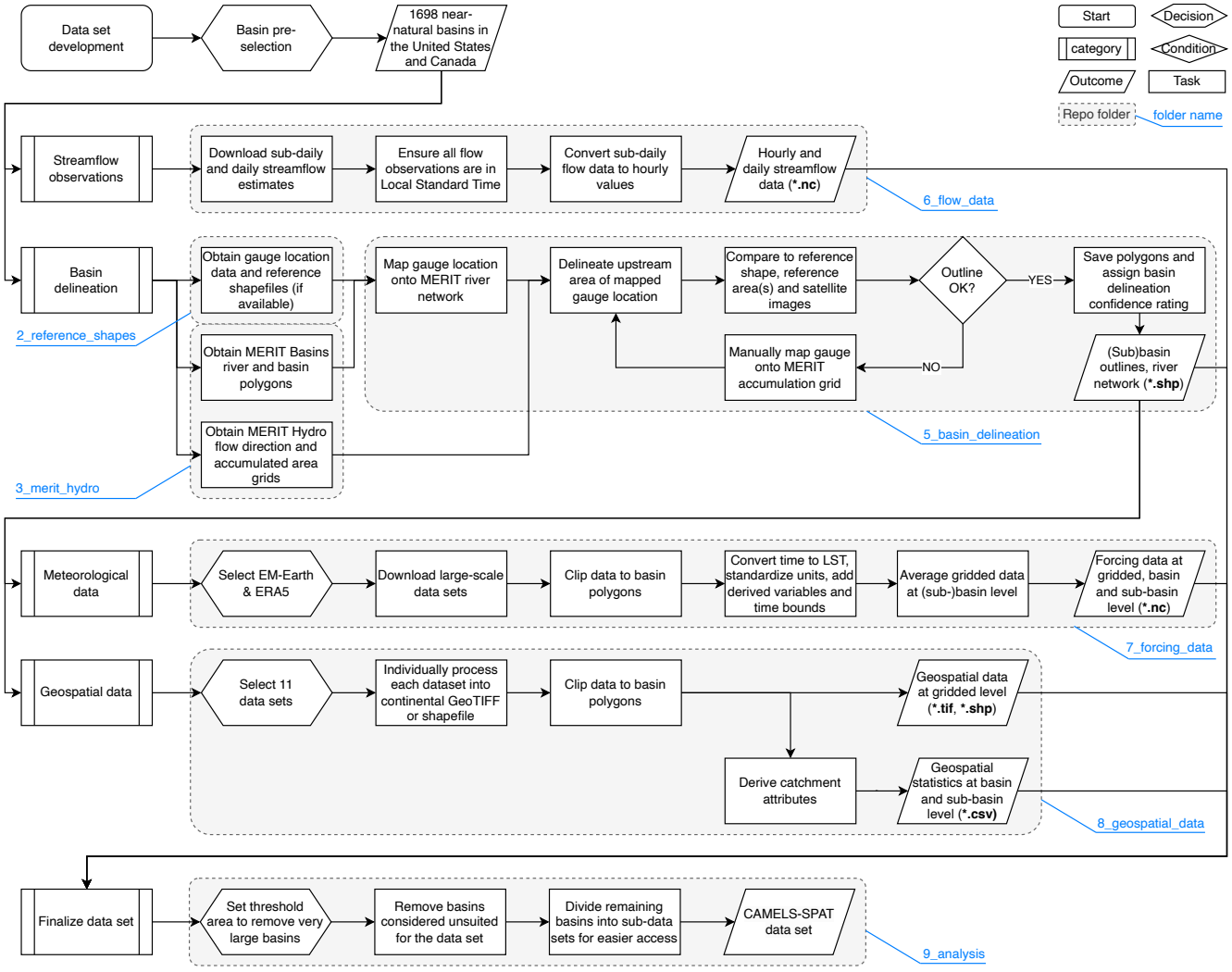

Figure 1. Overview of the CAMELS-SPAT workflow. Grey boxes and light blue call-outs indicate specific folders on the GitHub repository, where the necessary code to reproduce these steps can be found. **TO DO: drop number of data sets to 10 if we don't get permission to redistribute WorldClim**

streamflow observations at a sub-daily resolution. Process-based models are typically run at sub-daily time steps to more accurately simulate diurnal variation in processes such as evaporation, transpiration, sublimation and snow melt. In certain basins such diurnal variability is visible in the streamflow record, and such sub-daily observations are necessary to evaluate the appropriateness of process-based model equations. Daily data is by definition too coarse to distinguish such patterns.

2.1.2 Methods and outcomes

For basins in the United States, we rely on the basin selection made by Newman et al. (2015) that was used for the CAMELS data set (Addor et al., 2017a). This ensures some level of comparison between outcomes of studies using either CAMELS or CAMELS-SPAT is possible. We refer the reader to Section 2.1 in Newman et al. (2015) at all for a description of the criteria used to create this selection of 671 basins.

For basins in Canada, we start with the list of 1027 gauges included in the "Reference Hydrometric Basin Network" (RHBN, Environment and Climate Change Canada, 2020a, retrieved: 2022-08-18). These gauges have a minimum data availability of 20 years and minimal anthropogenic impacts as quantified by the presence of agriculture, built-up areas, and water management infrastructure, as well as population and road density. These criteria are comparable to those described in Newman et al. (2015). Note that agriculture presence in the Canadian prairie provinces (Alberta, Saskatchewan, Manitoba) and southern Ontario is substantial, and above the 10% area threshold used for the other provinces and territories (Pellerin and Nzokou Tanekou, 2020, p. 7). We attempt to address this by including various data products in CAMELS-SPAT that can be used to quantify or filter by the presence of agriculture.

Our initial basin selection thus includes 1698 basins. Various basins had to be removed due a lack of streamflow estimates or sub-daily data (see Section 2.3). We further removed several of the largest basins from the data set, under the assumption that any new insights that could be gained from these extremely large basins are minimal (especially given that these basins are under-gauged for their size) and do not outweigh the extra disk space and CPU time needed to run models in these regions. Our final selection consists of 1426 basins, divided into multiple sub-datasets intended to ease data access. For clarity, any outcomes shown in Sections 2.2 to 2.6 only show the final 1426 basins we have made publicly available, rather than the 1698 basins that are the outcome of this basin pre-selection step.

2.2 Basin delineation

2.2.1 Context

Hydrologic data sets such as this are conditional on having accurate basin outlines. Basin outlines are used to estimate a drainage basin's area, to crop meteorological and geospatial data from larger products to the area of interest, and to define the spatial extent of model configurations. Basin area estimates are also often used to convert the units of fluxes from volume-per-time to depth-per-time or vice versa (e.g. from $m^3 \cdot s^{-1}$ to $mm \cdot s^{-1}$). Using incorrect basin area estimates can lead to large conversion errors that propagate into any further analysis (McMillan et al., 2023).

The basin polygons provided as part of the CAMELS data (Newman et al., 2014; Addor et al., 2017b) are administrative boundaries. These polygons are not based on gauge locations, and the polygons thus tend to overestimate the basins' drainage areas. Estimated area errors are typically in the order of some percent (below 2% for approximately 70% of basins), but can be substantial (above 10% for some 8.5% of basins, with individual cases well above 100%). In addition, openly available polygons for the Canadian gauges did at the time of project initialization not fully cover all 1027 basins listed in the Reference Hydrometric Basin Network (Environment and Climate Change Canada, 2020b, retrieved: 2022-01-31).

To address both concerns, we delineated new basin outlines for all basins identified as potential candidates in Section 2.1. Our specific goals were to (1) identify the upstream area of each gauge, and (2) divide this upstream area into subbasin polygons of roughly equal size.

2.2.2 Method and outcomes

We obtained gauge metadata (location, name, reference areas, etc.), as well as reference basin outline polygons if these were available, for all gauges identified in the first step. For the US gauges, metadata and polygons showing each basin's outline were obtained from the CAMELS data set (Newman et al., 2014; Addor et al., 2017b). For the Canadian gauges, an initial download of the Reference Hydrometric Basin Network (RHBN) metadata is used to find gauge identifiers of those gauges that are in fact part of the RHBN updated in 2020. Further metadata (location, name) are then extracted from the HYDAT database (Environment and Climate Change Canada, 2010). Two different sets of reference polygons were available (Environment and Climate Change Canada, 2020b; Government of Canada, 2022, accessed: 2022-08-23, 2022-08-18, respectively), of which we preferentially used the newer polygons if these were available for our basins of interest.

To divide larger basins into smaller sub-basins we rely on the MERIT Basins data set (Lin et al., 2019). This data set contains vectorized river basins and river networks, derived from the MERIT Hydro data (Yamazaki et al., 2019). The mean (median) sub-basin size in the MERIT Basins data is 45.6 km^2 (36.8 km^2). We refer the reader to Lin et al. (2019) for further details. We also obtained the MERIT Hydro flow direction and accumulation grids (Yamazaki et al., 2019). The MERIT Hydro data is provided as gridded data in a regular longitude/latitude coordinate system (EPSG:4326). This is a common format (the meteorological data and many of the geospatial data sets we discuss in Sections 2.4 and 2.5 are also only available in EPSG:4326) and we adopt this as the standard in CAMELS-SPAT. The exception is area calculations and certain shapefile intersection operations, which are performed in the North America Albers Equal Area Conic projection (ESRI:102008).

The MERIT Basin network was derived independent from gauges and the sub-basins in this data set therefore do not align with gauge locations as reported by the United States Geological Survey and the Water Survey of Canada. For a given basin we thus needed to clip the most downstream sub-basin polygon to the gauge. We therefore first mapped the gauge locations onto the MERIT Hydro river network using automated techniques. This mapping is intended to guarantee that we start delineating the upstream basin from a pixel in the flow direction grid that is part of the main river (rather than the most downhill pixel of a single hillslope). However, there are various scenarios where automatic mapping is inaccurate and manual intervention is needed. We identified those cases through a combination of accuracy metrics (area comparison between new basin delineation and reported reference area(s), and percentage overlap between new basin delineation and reference polygon if any were available), and visual inspection of the new basin delineation, reference polygon, underlying MERIT Hydro data grids, and satellite images. If necessary, we manually defined a better outlet location to delineate the basin from and tracked this intervention in the CAMELS-SPAT metadata. We also assigned confidence ratings to our new basin polygons based on these quality assurance checks.

Figure 2 shows the resulting polygons for the 1426 basins that form the final CAMELS-SPAT data set, with colors indicating our confidence ratings. "Unknown" refers to cases where no confidence rating could be assigned, mainly due to lacking

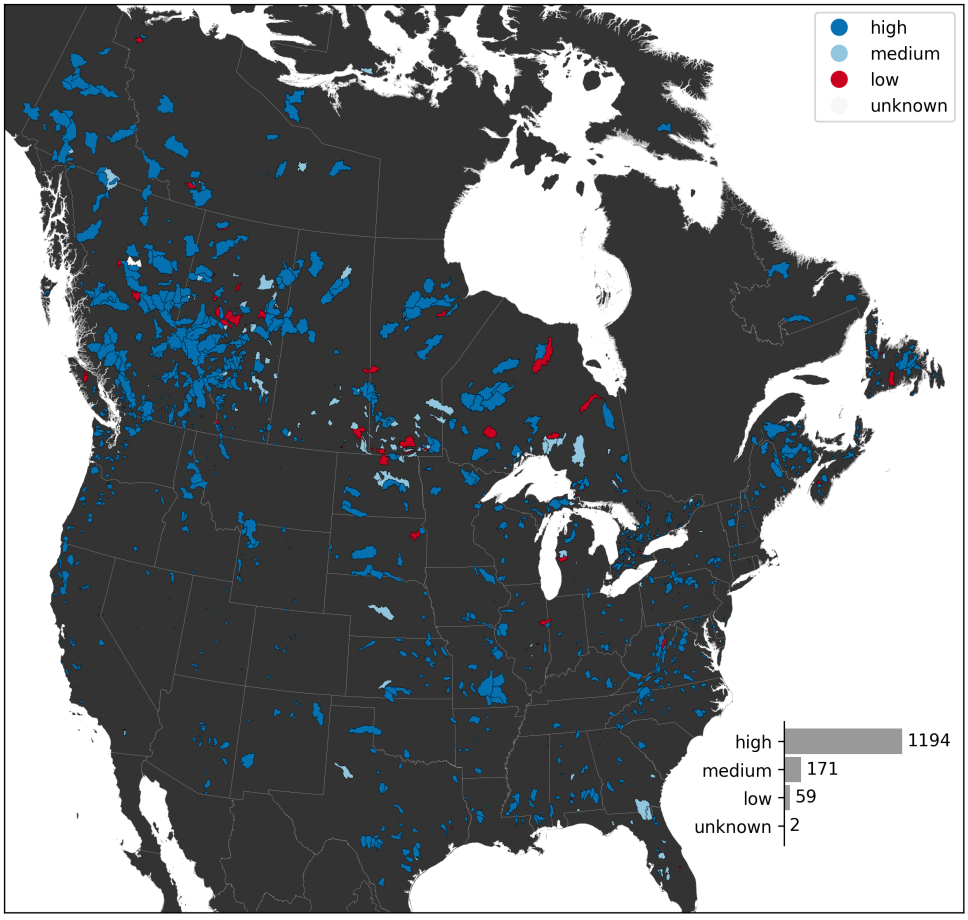

Figure 2. Location and delineation confidence of 1426 CAMELS-SPAT basins. Political boundaries by Commission for Environmental Cooperation (2022, accessed 2023-12-20)

.

reference polygons. “Low” ratings are assigned when evidence suggests that our basin delineations are inaccurate and we were unable to manually find a better outlet location that would lead to improved basin outlines. “Medium” ratings indicate that there are substantial differences between our new delineations and existing one and/or reference areas, but that it is difficult to decide if our new delineation or the reference(s) are inaccurate. “High” ratings are assigned when there is a clear match between our new polygons and the reference(s), or when evidence suggest out new delineations are more accurate than the reference(s). Detailed reasons for these ratings are tracked as part of the CAMELS-SPAT metadata. Medium and low confidence ratings occur primarily in regions with flat topography where finding the true outline of a drainage basin is difficult.

2.3 Streamflow observations

2.3.1 Context

Streamflow is a critical variable for many hydrologic studies. Streamflow estimates are typically provided as either instantaneous values (i.e., valid at a given point in time) or as averages over a given time interval.

The United States Geological Survey (USGS) typically collects instantaneous streamflow observations at 15- or 60-minute intervals. USGS also provides daily average values, computed from the instantaneous data from 00:00 to 24:00 Local Standard Time (LST; USGS, personal communication, 2023-06-20). Both instantaneous values and daily averages are publicly available.

The Water Survey of Canada (WSC) typically collects instantaneous streamflow observations at 5-minute intervals, and from these calculates daily averages that are reported in LST through the HYDAT database (WSC, personal communication, 2023-07-04). However, when instantaneous values are extracted through URL-based requests to the WSC webservice (https://wateroffice.ec.gc.ca/services/links_e.html), the time series are converted to UTC before being given to the user (Government of Canada, accessed: 2023-12-22). Instantaneous streamflow observations are publicly available for the period between present and minus 18 months. Daily average values are available for the full time period for which a gauge has been active.

Our goal with this project is to provide data useful for running and evaluating process-based hydrological models. We therefore include daily average streamflow values as available through USGS and WSC. We also include hourly average streamflow values when available, computed from the sub-daily instantaneous data available through both agencies. Note that all flow data, as well as meteorological forcing data, are included in the CAMELS-SPAT data set in Local Standard Time. LST of each gauge is tracked as part of the meta data.

2.3.2 Method and outcomes

For the gauges in the United States, daily average streamflow data and instantaneous (sub-daily) data can both be extracted through web requests (https://nwis.waterservices.usgs.gov/nwis/dv/ and https://nwis.waterservices.usgs.gov/nwis/iv/, respectively; accessed 2023-06-16). For the Canadian gauges, sub-daily data were extracted from the Environment and Climate Change Canada Web Service Links Interface (https://wateroffice.ec.gc.ca/services/links_e.html) on 2023-04-05. Daily data were extracted from the HYDAT database, version *20230505*. We excluded 4 gauges in the United States, as well as 180 Canadian gauges from the original 1697 preselected stations for lacking sub-daily data. We removed a further 13 Canadian gauges for lacking daily discharge values. Manual checks of these gauges through the WSC website (https://wateroffice.ec.gc.ca/search/historical_e.html) indicate that these stations are measuring water levels in lakes.

Daily average values for both countries are provided in LST. We updated the time indices for the sub-daily instantaneous values to match. For the gauges in the USA, this meant shifting the time series by 1 hour for time steps that were provided in local daylight saving time, for gauges in states where daylight saving time is observed. For the Canadian gauges, this meant shifting the entire time series for each gauge by the offset needed to convert UTC to LST. We then set any negative streamflow values to zero, and used a mass-conserving averaging approach to turn instantaneous flow data into hourly averages (see

Appendix A). We specified the condition that every hourly average must be based on at least one observation during that time window. Hours for which no data observations were available were set to Not-a-Number (NaN).

Note the critical assumption that we calculated the average hourly flows reported value at the full hour (e.g., 12:00) using a forward-looking window (i.e., in this case the value at 12:00 is the average during the time window 12:00-13:00). This matches the daily flows, which are provided under the same assumption by USGS and WSC (e.g., the Jan-1 2000 value is calculated from data between 00:00 Jan-1 and 24:00 Jan-1; USGS, personal communication, 2023-06-20; WSC, personal communication, 2023-06-26). This information is also stored in the *time_bnds* (time bounds) variable available in the provided NetCDF files.

Daily and sub-daily observations were originally provided in text-based formats. We converted these to NetCDF4 formats, to ensure consistency between gauges in the two countries and to track metadata in a more accessible way (compared to storing the metadata in separate files or headers in text files). For both USGS and WSC data we retained the quality flags that accompany the data and stored these next to the data in the NetCDF files.

Figure 3 shows aggregated flow data availability for the 1426 catchments included in the CAMELS-SPAT data set. Hourly flow data comes in two distinct categories: short (< 2 years) records for the Canadian gauges and much longer records for gauges located in the United States. This is a consequence of Water Survey of Canada's policy to make high-resolution gauge data only publicly available for a short historical period. Missing data for these shorter records are however typically low (see also Fig. B1). For approximately 80% of gauges, missing hourly observations account for up to 10% of record length. Data may be missing for up to 40% of the record for most remaining gauges, with a handful of gauges having extremely large data gaps. Daily data record lengths are similar for Canadian and United States gauges. Missing values are relatively rare (<10% for up to 1350 out of our 1426 gauges), though can be substantial (up to 80 to 95% for the remaining gauges; see Fig. B1). The period with the greatest overlap of data records is 1990-2020; hourly observations are available for only a handful of gauges before this time.

2.4 Forcing data

2.4.1 Context

Meteorological forcing data in existing data sets is typically provided as catchment-averaged (lumped) daily data, and tends to be limited to precipitation, temperature and potential evapotranspiration variables (Table 1). While a large swathe of the more conceptual models can be run with just precipitation, temperature and potential evapotranspiration inputs (see e.g., Knoben et al., 2019; Trotter et al., 2022) at coarse temporal and spatial resolutions, such choices are typically insufficient for the more complex hydrologic models.

Table 2 shows a brief overview of meteorological data requirements for a selection of process-based hydrological models. Typical variables include (1) precipitation, (2) air temperature, (3) radiation terms, often distinguishing between shortwave and longwave radiation, (4) air pressure, (5) humidity, and (6) wind speed. Further differences between simpler and more complex models can be found in their typical temporal and spatial resolutions. Process-based models are typically run at sub-daily time steps to capture diurnal variations in processes such as evaporation, transpiration and snow melt. Such models can also be run

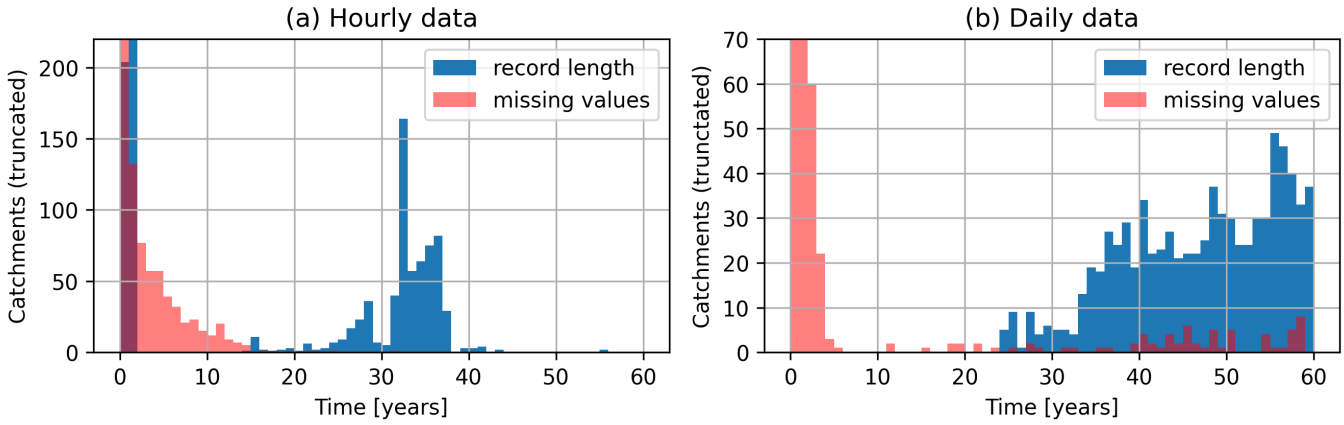

Figure 3. Flow data availability for gauges included in CAMELS-SPAT. Record length refers to the period between the first publicly available flow record for a given station, and its last. Missing values occur within this record period and are given here in the same units as the record length itself. Note that both y-axes are truncated: in (a), *missing values* has a count of 913 for *time* is [0,1], and *record length* has a count of 560 for *time* = [1,2]; in (b), *missing values* has counts of 1156 and 112 for *time* is [0,1] and [1,2], respectively.

in fully-distributed (i.e., gridded) or semi-distributed (i.e., on sub-basins) fashions to account for spatial heterogeneity within the larger basin.

It is clear from Table 2 that it is impossible to define a small set of forcing variables that would allow the use of a large number of process-based hydrologic models. We therefore decided to include a broad selection of meteorological variables, accepting that this comes at the cost of extra disk space. We provide these variables at hourly time steps, at their native gridded resolution as well as averaged at the sub-basin level. To facilitate the use of the broadest range of modeling tools we also include time series of potential evaporation (see footnote in Table 3) and forcing variables aggregated at the whole basin level.

Table 2. Meteorological data needs for CATFLOW (Maurer and Zehe, 2007), CHM (Marsh et al., 2020), CHRM (Pomeroy et al., 2007), ES-CROC (Lafaysse et al., 2017), HYPE (SMHI, 2022), MESH (Mekonnen and Brauner, 2020), Noah LSM (Mitchell et al., 2005), PARFLOW (Maxwell et al., 2019), PIHM (PIHM team, 2007), SUMMA (Clark et al., 2015a, b; Nijssen, 2017), VIC (Liang et al., 1994; Hamman et al., 2018) and WaSIM (Schulla, 2021). Models are listed alphabetically. Optional inputs indicated with *.

Variable	CATFLOW	CHM	CHRM	ES-CROC	HYPE	MESH
Precipitation	$[m\,t^{-1}]$	$[mm\,t^{-1}]$	$[mm\,t^{-1}]$	$[kg\,m^{-2}\,s^{-1}]$	$[mm\,t^{-1}]$	$[kg\,m^{-2}\,s^{-1}]$
Downward shortwave radiation	$[W\,m^{-2}]$	$[W\,m^{-2}]$		$[W\,m^{-2}]$	$[MJ\,m^{-2}\,d^{-1}]$*	$[W\,m^{-2}]$
Downward longwave radiation		$[W\,m^{-2}]$		$[W\,m^{-2}]$		$[W\,m^{-2}]$
Air temperature	$[C]$	$[C]$	$[C]$	$[K]$	$[C]$	$[K]$
Air pressure				$[Pa]$		$[Pa]$
Specific humidity						$[kg\,kg^{-1}]$
Wind speed (U-direction)					$[m\,s^{-1}]$*	
Wind speed (V-direction)					$[m\,s^{-1}]$*	
Sunshine duration						
Reflected shortwave radiation			$[W\,m^{-2}]$			
Net radiation	$[W\,m^{-2}]$		$[W\,m^{-2}]$			
Vapor pressure						
Relative humidity	$[\%]$	$[\%]$	$[\%]$	$[\%]$	$[-]$*	
Wind speed (mean)	$[m\,s^{-1}]$	$[m\,s^{-1}]$	$[m\,s^{-1}]$	$[m\,s^{-1}]$	$[m\,s^{-1}]$	$[m\,s^{-1}]$
Wind direction	$[degrees]$	$[degrees]$				

Variable	Noah LSM	PARFLOW	PIHM	SUMMA	VIC	WaSIM
Precipitation	$[inch\,30min^{-1}]$	$[mm\,s^{-1}]$	[?]	$[kg\,m^{-2}\,s^{-1}]$	$[mm\,t^{-1}]$	$[mm]$
Downward shortwave radiation	$[W\,m^{-2}]$	$[W\,m^{-2}]$	[?]	$[W\,m^{-2}]$	$[W\,m^{-2}]$	$[Wh\,m^{-2}]$
Downward longwave radiation	$[W\,m^{-2}]$	$[W\,m^{-2}]$		$[W\,m^{-2}]$	$[W\,m^{-2}]$	
Air temperature	$[C]$	$[K]$	[?]	$[K]$	$[C]$	$[C]$
Air pressure	$[mbar]$	$[Pa]$		$[Pa]$	$[kPa]$	
Specific humidity		$[kg\,kg^{-1}]$		$[g\,g^{-1}]$		
Wind speed (U-direction)		$[m\,s^{-1}]$				
Wind speed (V-direction)		$[m\,s^{-1}]$				
Sunshine duration						$[-]$
Reflected shortwave radiation						
Net radiation						
Vapor pressure			[?]		$[kPa]$	
Relative humidity	$[-]$		[?]			$[-]$
Wind speed (mean)	$[m\,s^{-1}]$		[?]	$[m\,s^{-1}]$	$[m\,s^{-1}]$	$[m\,s^{-1}]$
Wind direction						

2.4.2 Methods and outcomes

For internal consistency of the CAMELS-SPAT data, we obtained meteorological variables from forcing products with global coverage, rather than relying on national products. We primarily use ERA5 data (Hersbach et al., 2020, available at 0.25° resolution), complemented with high-resolution precipitation and temperature data from deterministic EM-Earth (Tang et al., 2022b, available at 0.25° resolution). Compared to EM-Earth, ERA5 provides a wide range of variables relevant for hydrologic modeling, but initial findings suggest that the two variables EM-Earth does provide lead to better modeling results for our area of interest than those same variables in ERA5 do (Rakovec et al., 2023). However, note that the EM-Earth has a fixed temporal coverage of 1950-2019, whereas our selected gauges have data beyond 2019. EM-Earth does thus in many cases not cover the full observation period for each gauge.

Table 3 shows an overview of forcing variables available as time series in the CAMELS-SPAT data set. Compared to Table 2, we provide net radiation terms at the surface separated into net shortwave and net longwave terms, and do not provide a summed net radiation component nor a reflected shortwave variable. Either can be easily derived from the provided net shortwave and longwave component (see Hogan (2015), but also footnote 2 in Table 3). We also do not provide sunshine duration because this is not a independent variable available in ERA5: it is derived from downward shortwave radiation using a threshold of 120 $W\ m^{-2}$ (Hogan, 2015). Hogan (2015) also cautions against its use, because its definition in ERA5 differs from standard. We complement ERA5 data with various additional variables derived from the downloaded data in cases where we judged the processing to be too cumbersome to pass down to the user (i.e., vapor pressure, relative humidity, wind direction), or the variable seemed to be of general interest (i.e., mean wind speed). Equations of derived data are provided in Appendix C. While this list of variables is unlikely to completely cover all models' data needs, it will provide a reasonable starting point for a large number of models.

We have chosen to keep the ERA5 data in its original units, and convert the EM-Earth units to match (i.e., precipitation from $[mm\ h^{-1}]$ to $[kg\ m^{-2}\ s^{-1}]$ assuming a water density of 1000 $kg\ m^{-3}$, and temperature from $[C]$ to $[K]$). We also retained the original variable names so that users may easily refer to the existing documentation of both ERA5 and EM-Earth if needed. Data are provided for the full time period covered by the observational record of each individual gauge when possible, including time steps for which streamflow data are missing.

In addition to time series, we also included various time-invariant ERA5 data. Most notably this includes geopotential, which can be used to approximate the mean elevation of each ERA5 grid point, and is thus useful when applying elevation-dependent lapse rates to any forcing variables. For all variables, metadata (descriptions, units, derivations if applicable) are stored as variable attributes in the NetCDF files.

We provide the forcing data at three different spatial aggregation levels: (1) as gridded values at the native resolution of each data set, clipped to the basin outline; (2) aggregated at the sub-basin level; (3) aggregated at the basin level (i.e., the level at which most of the data sets listed in Table 1 provide data). Averaging of the gridded data to (sub-)basin polygons was done with the EASYMORE toolbox (Gharari et al., 2023a).

Both ERA5 and deterministic EM-Earth provide data at hourly resolution, in Coordinated Universal Time (UTC). For ease of use we process these time indices to be in each gauge's Local Standard Time (LST) instead. Note that ERA5 mean variables use *period-ending* time stamps, meaning that, for example, the average precipitation rate at time 12:00 is the average rate over the interval 11:00-12:00 (European Centre for Medium-range Weather Forecasting, 2023b, Section: "Mean rates/fluxes and accumulations"). EM-Earth's precipitation variable instead uses *period-beginning* time stamps, meaning that, for example, the average precipitation rate at time 12:00 is the average rate over the interval 12:00-13:00 (G. Tang, personal communication, 2024). Instantaneous variables are valid at the specific point in time given by the timestamp (European Centre for Medium-range Weather Forecasting, 2023c). The column *Validity* in Table 3 indicates how variables should be interpreted.

2.5 Geospatial data

2.5.1 Context

Geospatial data in existing data set covers four broad categories: (1) meteorology (as time series and derived summary statistics), (2) vegetation and land use; (3) topography; (4) soil and geology. Such data sets are typically provided as maps in their original formats, but tend to be presented as spatial statistics (mean, mode, etc.) in various large-sample data sets. These statistical summaries of the original data can be helpful to succinctly characterize a location's hydroclimatic conditions and support classification efforts. For modeling purposes, models such as Noah-LSM (Niu et al., 2011) and SUMMA (Clark et al., 2015a, b) rely on vegetation and soil classes to provide initial values for a number of land use and soil parameters. More generally, model parameters can in certain cases be derived from geospatial data. Examples can include the height of the vegetation canopy in the vertical direction, or the fraction open water in the horizontal direction.

Existing large-sample data sets cover a wide variety of attributes. An informal analysis of the CAMELS data sets listed in Table 1 shows that these data sets together contain close to 300 different attributes, though any given individual data set contains no more than 50 to slightly over a 100 of those. This lack of uniformity is compounded by a lack of unified terminology. This is in line with findings by (Tarasova et al., 2023), who analyze how 742 journal articles describe the hydroclimatic conditions of their study areas. They find that authors use a wide variety of attributes with only occasional verification of their attributes' usefulness. Relevant for our work, and in line with a cursory overview of attributes provided by the data sets listed in Table 1, they also find that the existing literature only rarely uses catchment descriptors that attempt to quantify the range a particular variable may cover in a given catchment. As a way forward, they recommend "hypothesis-oriented selection of catchment descriptors" on a per-study basis.

Given (1) the difficulty to define a set of catchment attributes consistent with other available data sets, (2) the subjectivity concerns outlined by Tarasova et al. (2023), and (3) our goal to facilitate a wide range of modeling studies, we have chosen to provide only maps of geospatial data as part of CAMELS-SPAT. We have not attempted to summarize these maps into relevant statistical descriptors on a per-(sub-)basin basis for fear of our selection becoming the "default" set of attributes for studies using the CAMELS-SPAT data. While this will require more effort on the user's side to determine appropriate statistics for their study and process our provided maps, we believe that this approach will lead to more defensible science in the long term.

Table 3. CAMELS-SPAT meteorological variables. Note that ERA5 and EM-Earth data are stored in separate NetCDF files. Derived variables are described in more detail in Appendix C. "Validity" indicates how variable time steps must be interpreted. Period-beginning variables are mean rates valid over a time interval between the current time step and the next (e.g. the value at 12:00 is valid for the window 12:00-13:00); Period-ending variables are mean rates valid over a time interval between the previous time step and the current (e.g. the value at 12:00 is valid for the window 11:00-12:00); instantaneous variables are valid at a specific point in time (e.g. the value at 12:00 is valid at 12:00.

Source	Variable	Units	Name in NetCDF	Validity
EM-Earth	Mean total precipitation rate	$[kg\ m^{-2}\ s^{-1}]$	*prcp*	Period-beginning[5]
EM-Earth	Temperature	$[K]$	*tmean*	Instantaneous[5]
ERA5	Mean total precipitation rate	$[kg\ m^{-2}\ s^{-1}]$	*mtpr*	Period-ending[6]
ERA5	Mean downward shortwave radiation at the surface	$[W\ m^{-2}]$	*msdwswrf*	Period-ending[6]
ERA5	Mean downward longwave radiation at the surface	$[W\ m^{-2}]$	*msdwlwrf*	Period-ending[7]
ERA5	Net shortwave radiation at the surface[1]	$[W\ m^{-2}]$	*msnswrf*	Period-ending[6]
ERA5	Net longwave radiation at the surface[1]	$[W\ m^{-2}]$	*msnlwrf*	Period-ending[6]
ERA5	Mean potential evaporation rate[2]	$[kg\ m^{-2}\ s^{-1}]$	*mper*	Period-ending[6]
ERA5	Temperature	$[K]$	*t*	Instantaneous[7]
ERA5	U-component of wind	$[m\ s^{-1}]$	*u*	Instantaneous[7]
ERA5	V-component of wind	$[m\ s^{-1}]$	*v*	Instantaneous[7]
ERA5	Specific humidity	$[kg\ kg^{-1}]$	*q*	Instantaneous[7]
ERA5	Surface pressure	$[Pa]$	*sp*	Instantaneous[8]
Derived	Vapor pressure	$[kPa]$	*e*	Instantaneous
Derived	Relative humidity	$[kPa\ kPa^{-1}]$	*rh*	Instantaneous
Derived	Wind speed (mean)	$[m\ s^{-1}]$	*w*	Instantaneous
Derived	Wind direction[3,4]	$[degrees]$	*phi*	Instantaneous

[1]Note that these net radiation terms are based on interactions between the atmospheric and land surface components of the ERA5 modeling chain, and should thus only be used carefully as model input to prevent cases where the user's model duplicates processes already accounted for by the ERA5 models.

[2] We retain the use of ECMWF's naming here rather than the common "potential evapotranspiration". Assumptions underlying this variable are described here: https://codes.ecmwf.int/grib/param-db/?id=228251 (last access: 2024-01-01). Note that we provide the equivalent variable as a mean rate as part of the CAMELS-SPAT data, but the url for that variable lacks a clear description: https://codes.ecmwf.int/grib/param-db/?id=235070 (last access: 2024-01-01).

[3]We derived vapor pressure, relative humidity and mean wind speed before averaging the gridded data onto (sub-)basins, but this is not easily possible for wind direction. Instead, we calculate wind direction separately for the gridded, semi-distributed and lumped cases from u- and v-components after (sub-)basin averages of these variables were created.

[4] Note we use the meteorological wind direction as defined by ECMWF (European Centre for Medium-range Weather Forecasting, 2023a): wind direction in this case indicate the direction the wind *comes from*, not where it goes.

[5] G. Tang, personal communication, 2024.

[6] See: https://confluence.ecmwf.int/pages/viewpage.action?pageId=82870405#ERA5:datadocumentation-Table4 (last access: 2024-01-03)

[7] See: https://confluence.ecmwf.int/pages/viewpage.action?pageId=82870405#ERA5:datadocumentation-Table9 (last access: 2024-01-03)

[8] See: https://confluence.ecmwf.int/pages/viewpage.action?pageId=82870405#ERA5:datadocumentation-Table2 (last access: 2024-01-03)

2.5.2 Methods and outcomes

For internal consistency of the CAMELS-SPAT data, we selected various geospatial data sets that cover at least the United States and Canada. The specific processing steps vary, but in general for each data set processing involved downloading the data at continental or larger scales and clipping the data to the basin polygons. We also ensured all data are provided in a regular latitude/longitude coordinate system if they weren't already. Figure 4 provides an overview of the geospatial data layers, using a single basin as an example.

Long-term monthly means of several climate variables can be obtained from the WorldClim data set (Fick and Hijmans, 2017). The advantage over calculating these means from the ERA5 data is WorldClim's much higher spatial resolution. WorldClim's data license does not allow redistribution of their data, so we encourage interested users to use the code available on our GitHub repository to obtain and preprocess this data set. Available variables are long-term means computed from 30 years each, showing minimum, mean and maximum monthly temperature, as well as monthly precipitation, solar radiation, wind speed and water vapor pressure. These variables may be useful for water balance investigations, bias correction of ERA5 and EM-Earth, and climate classification efforts.

Process-based hydrological models typically include explicit representations of vegetation cover in a catchment. CAMELS-SPAT includes two data sets from which vegetation parameters may be derived. First, we included time series of Leaf Area Index (LAI) observations, derived from MODIS satellite observations (Myneni et al., 2021, MCD15A2H.061). These observations are available at an 8-day temporal resolution and cover the period 2002-07-04 to 2023-10-08. Certain models may be able to ingest these maps directly, or typical seasonal LAI patterns may be derived from therm. In addition, we included estimates of forest height in 2000 and 2020 (Potapov et al., 2021, part of the Global Land Cover and Land Use Change, 2000-2020 data).

To further assist parametrization and classification efforts, we included three different products related to land cover and land use. First, the Landsat-Derived Global Rainfed and Irrigated-Cropland Product (LGRIP30, Thenkabail et al., 2021; Teluguntla et al., 2023) can be used to estimate the magnitude and type of agriculture practiced in each basin. Second, we include per basin a map of International Geosphere–Biosphere Programme (IGBP) land classes, derived from MODIS satellite observations (Friedl and Sulla-Menashe, 2022). Land use type is a mandatory input for various process-based hydrological models to activate certain process modules or estimate parameter values. Third, we include high-resolution Global Land Cover and Land Use 2019 maps (Hansen et al., 2022). This is very high-resolution data derived from Landsat satellite observations, used to classify the landscape a few broad categories (inland water, permanent snow and ice, cropland, built-up, terra firma and wetlands) with several of these consisting of subclasses based on build-up area extent, and vegetation extent and height.

We include cutouts of the HydroLAKES data (Messager et al., 2016) to quantify the extent, type and volumes of open water bodies in each basin. This data can be used to estimate each catchment's open water area, retention volumes and parametrization of reservoir and lakes modules in hydrologic and/or routing models.

The MERIT Hydro Digital Elevation Model (DEM) used for basin delineation (Yamazaki et al., 2019) is also part of the maps provided for each catchment. Additional variables such as slope and aspect may be derived from this map, as well as elevation bands. This data can also be useful in combination with the ERA5 time-invariant geopotential variable (see Section

2.4) to apply lapse rates to meteorologic variables. **Maybe I should put those slope and aspect maps in already - it makes sense**.

Finally, we provide maps from three different data sets to characterize each catchment's subsurface. First, SOILGRIDS 2.0 (Poggio et al., 2021) provides estimates of various soil properties (bulk density, percentage coarse fragments, organic carbon content, and sand, silt and clay percentages) at six different depths (0-5cm, 5-15cm, 15-30cm, 30-60cm, 60-100cm, 100-200cm). From these, soil classification is possible. These maps are given for mean values, but also for 0.05^{th}, 50^{th} and 95^{th} percentiles and an uncertainty estimate. However, SOILGRIDS data are estimated for depths up to 2 meters everywhere, without taking into account the actual depth to bedrock of any location. Thus, second, we included maps from the Pelletier soil database (Pelletier et al., 2016b, a). These distinguish between uplands, valley bottoms and lowlands and provide estimates of the depths of soil, intact regolith, and sedimentary deposits above unweathered bedrock. These variables may be used to set more realistic soil depths in models compared to a spatially uniform depth. Third, we include cut-outs from the GLHYMPS data (Gleeson et al., 2014; Gleeson, 2018) as polygons. Contained as attributes are estimates of permeability and porosity, which may be used to parametrize models (for example, by deriving hydraulic conductivity which is a common model parameter).

2.6 Data division

2.6.1 Context

As noted in the introduction, one benefit of large-sample over large-scale data sets is their ease-of-use (in terms of disk space and computational effort). In terms of area, CAMELS-SPAT covers a fairly small area compared to the land extent of the North American continent (Figure 2. However, in terms of data set size our original selection of 1697 stations required approximately 8 TB of disk space. This should be no concern in High Performance Computing environments, but such requirements (still) outstrip easily available storage on personal computing devices.

2.6.2 Methods and outcomes

Figure 5 shows an overview of how data set size changes as a function of area thresholds used to determine which basins are included in the data set. Disk space requirements are highly non-linear. If all 1697 basins are included, the total disk space requirements are approximately 8 TB. Based on this diagram, we decide to exclude any basins with an area larger than $10^4\ km^2$ from the final data set distribution. This reduces the number of gauges by 113, but halves the disk space requirement of the total data set.

For convenience, we divided the final selection of 1426 gauges into various subsets. First, we divided the data set into three categories of headwater, meso-scale and macro-scale basins. Headwater basins are defined as catchments with only a single sub-basin in our delineation. Meso-scale basins are basins that are not headwaters and below a total area of $10^3\ km^2$, and macro-scale basins are those with areas between $10^3\ km^2$ and $10^4\ km^2$. Headwater basins account for 307 out of 1426 total, at approximately 430 GB disk space. 724 basins fall in our meso-scale category (mean area of approximately 400 km^2, with

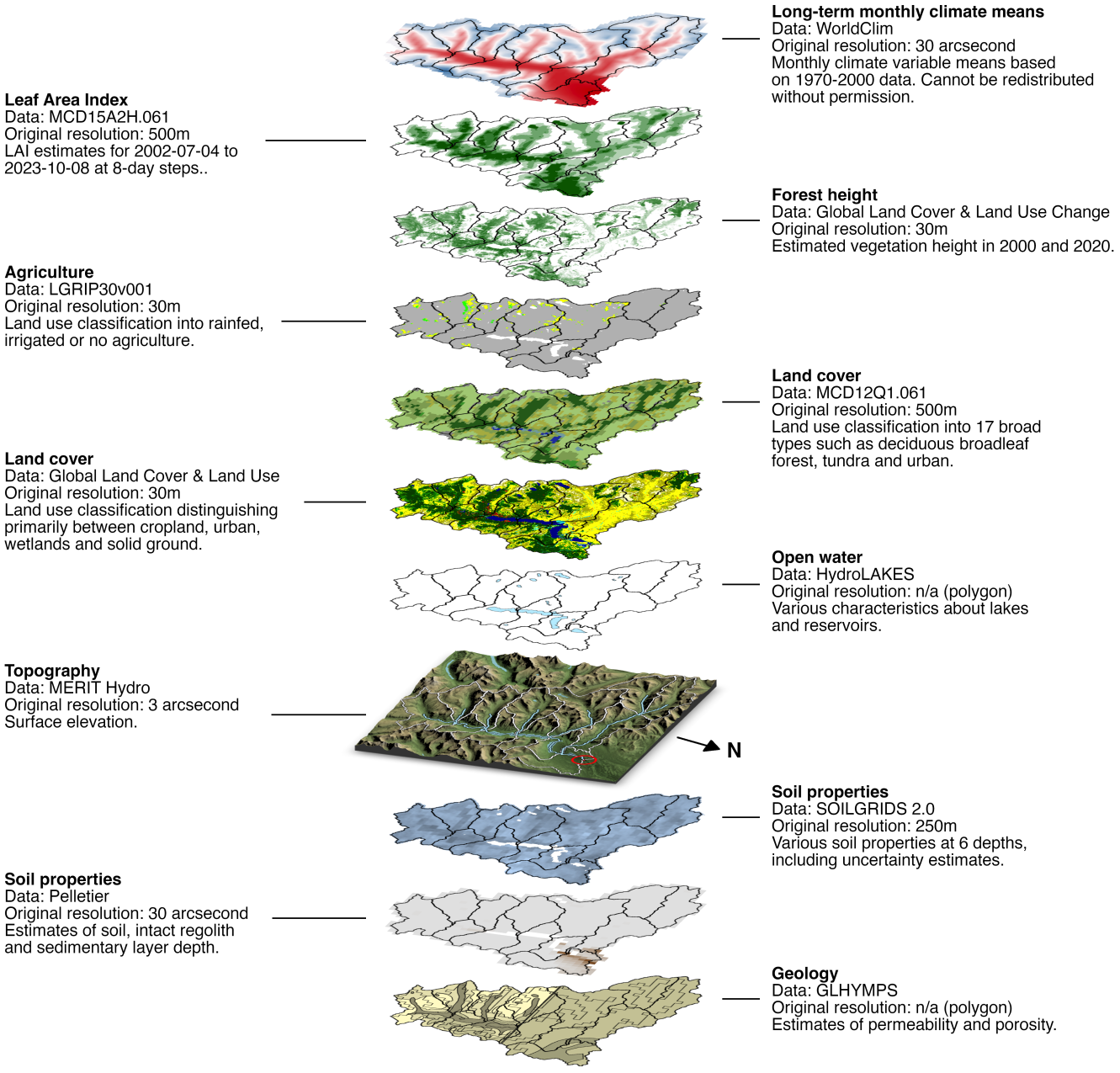

Figure 4. Overview of geospatial maps provided for each catchment in the CAMELS-SPAT data set, using a transboundary basin as an example (Canadian gauge ID: *05AD003*).

on average 9 sub-basins), at approximately 1400 GB disk space. Macro-scale basins (mean area $\approx 3100\ km^2$, on average 66 sub-basins) are the remaining 446 basins, at 1600 GB disk space.

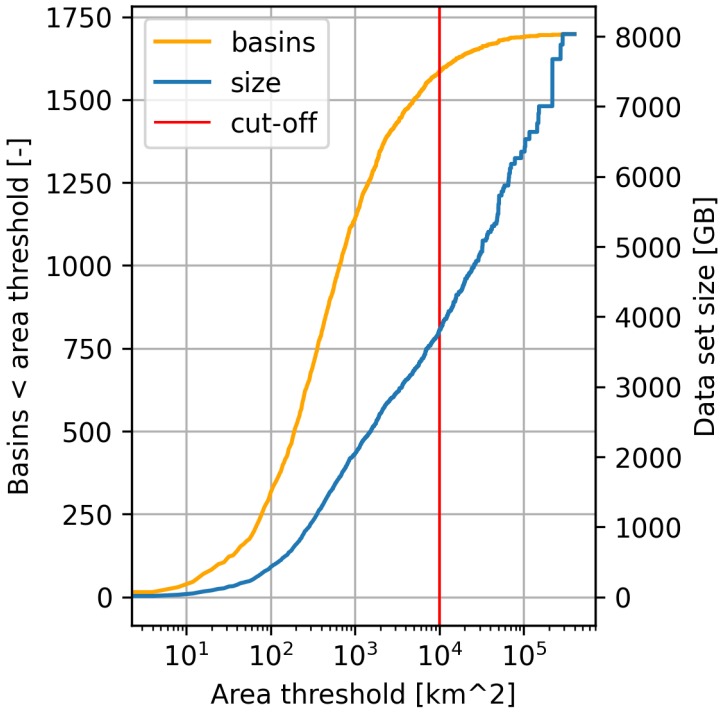

Figure 5. Overview of data set size (number of basins on the left y-axis, disk space requirements on the right y-axis) as a function of area thresholds below which basins are included in the counts. On the left, only the smallest basins are included and disk space requirements are consequentially small. On the right, all basins are included resulting in large disk space requirements.

3 Discussion

3.1 Recommendations for data providers

3.1.1 Dimension boundary information in publicly available data

In Sections 2.3 and 2.4 we describe the processing of streamflow observations and meteorological data, respectively. One challenge here is determining the representativeness (or validity) of data values in time and space. Data can be instantaneous (i.e., valid at a specific point in time) or time-averaged (i.e., valid over a specific time window), and treating one as the other leads to incorrect estimates of fluxes and thus state changes in the system (see e.g., Appendix A). The same concern applies to space: values may be representative for a specific point, or averaged over a given region. Accounting for these differences is not always straightforward, in particular because information about the spatial and temporal validity of publicly available data is not always easily available and may require informal inquiries to obtain. This hampers the correct application and interpretation of data, and can lead to easily preventable biases in analyses and modeling efforts.

A simple solution is provided by the NetCDF Climate and Forecast (CF) Metadata Conventions (Eaton et al., 2023, see Section 7). These conventions describe the specification of bounds for coordinate variables (i.e., dimensions such as latitude, longitude and time) that indicate between which coordinate values a given data value is considered valid. Specific examples for spatial, gridded data can be found in Section 7.1 in Eaton et al. (2023); time bounds are discussed in Examples 7.5 and 7.6. The CF conventions are designed for NetCDF files but the principle of specifying dimension bounds in time and space, between which data values are valid, is widely applicable. We strongly recommend that including these bounds as part of data distributions becomes standard practice.

3.1.2 Sub-daily flow data derivations

Process-based models can be useful for long-term water assessments, provided that they are parametrized well and that the theoretical underpinnings of the model are valid (e.g., Kirchner, 2006; Clark et al., 2016). In the case of process-based models, assessing a model's physical realism requires observations at sub-daily resolution. In CAMELS-SPAT we therefore construct hourly streamflow series from time series of instantaneous streamflow observations that are publicly available. However, the phrase "streamflow observations" (though common) is somewhat misleading: in almost all cases the observations are of water levels and streamflow values are estimated for a given water level with rating curves. Especially at high observation frequencies these water levels may be subject to random fluctuations unrelated to streamflow magnitude (e.g., due to wind or small eddies), which will translate into streamflow estimates affected by this noise. A cleaner approach would be to find the average hourly water level, and estimate the average hourly flow from this through the station's rating curve. Development and maintenance of rating curves is complex however and rating curves tend to change through time (see for example the description of WSC's procedures in Gharari et al., 2023b). Computing robust sub-daily streamflow estimates will be easier at institutional levels (not least because it requires access to the rating curves) and we express the hope that this may become standard practice.

Martyn, Al: suggestions for a better phrase or thought to end this paragraph on are very welcome

3.2 Guidelines for practical use

Here we outline various considerations that may be useful to readers. Our goal with these is to set expectations for data set use, and highlight potential pitfalls that may not be immediately obvious.

3.2.1 Utilization of streamflow data quality flags

We retained streamflow observation quality flags provided by the USGS and WSC during processing and stored these in the same NetCDF files as the streamflow observations themselves. These flags indicate conditions affecting the streamflow measurement, such as the presence of river ice, backwater effects, water levels below sensor level, or equipment malfunction. These conditions suggest that streamflow data at these time steps are inaccurate and this may affect analyses that use these data. For example, it is known that errors at individual time steps may have disproportionate effects on aggregated efficiency scores

that are used in modeling (e.g., Newman et al., 2015; Clark et al., 2021). Excluding streamflow observations from efficiency score calculations based on data quality flags is a possible way to limit the impacts of known erroneous streamflow values.

3.2.2 Spatial validity of meteorological forcing data

CAMELS-SPAT contains meteorological data from both ERA5 and EM-Earth, at their original gridded resolution as well as averaged at the basin and subbasin level. During this averaging process we assumed that values provided at specific coordinates are valid for a grid cell around this point. This is a simplistic approach but it is somewhat difficult to justify more elaborate assumptions (such as some form of interpolation), because in reality the change of meteorological variables in space would be dependent on local topography at scales smaller than the typical forcing data grid cell. Interpolation methods may yield more realistic subbasin and basin averaged values, but it is beyond the scope of this paper to investigate these. Alternatively, application of lapse rates using the available elevation data and ERA5's time-invariant geopotential may go some way to address this.

3.2.3 Internal consistency of meteorological variables

Martyn, Al: I tried unsuccessfully to turn the bullet points below into prose. I only have a rudimentary understanding of meteorology and it's making this a little hard to write. Before I spent more time on this, (a) can you tell me if this line of reasoning is even correct, (b) and if so, worth worrying about?

- camels-spat has ERA5 and EM-Earth
- ERA5 has a collection of variables that can serve as model inputs, EM-Earth only has 2
- for certain purposes it may be helpful to complement EM-Earth data with something else but this must be done with caution
- if EM-Earth temperature and/or precip values are different than ERA5's, this suggest that in EM-Earth's version of reality the weather system behaved differently than in ERA5's
- mashing variables from different data sets together does not necessarily lead to a selection of variables per timestep that are internally consistent
- hard to estimate the impact of this, hence this warning

3.2.4 Handling the "graveyard of hydrological models"

Model performance across the United States is known to change regionally, where model performance is at its worst in the drier central regions (e.g., Newman et al., 2015; Towler et al., 2023). In CAMELS-SPAT we compound this problem by including basins from an area colloquially known as "the graveyard of hydrological models" (e.g., Muhammad et al., 2019; Budhathoki et al., 2020; Ahmed et al., 2023). This region, also known as the "Prairie-Pothole Region", covers parts of southern

Alberta, Saskatchewan, Manitoba, North Dakota, South Dakota, Minnesota and Iowa. The landscape here is relatively young on geological time scales and large parts of it have not yet eroded into traditional river networks. Surface depressions are common and typically not connected to the stream network, except through very slow groundwater drainage and the occasional fill-and-spill event (Hayashi et al., 2016; Clark and Shook, 2022). In the basins we provide as part of the CAMELS-SPAT data, all subbasins are connected to the stream network. However, surface depressions below the resolution of the MERIT DEM are common and will affect hydrologic behaviour in these (sub)basins. We recommend that users account for these potholes in their analyses and modeling efforts, possibly through the use of stand-alone models or post-processing tools (e.g., Clark and Shook, 2022), or by adapting existing models with an appropriate landscape module (e.g., Ahmed et al., 2023), or to adjust their expectations about model performance accordingly.

3.2.5 Extension of catchment attributes

We derived various catchment attributes for the basins in CAMELS-SPAT, for ease of use and comparison with existing data sets.

Hypothesis-based catchment attributes (cross-correlations, independent information), new attributes can be derived
NOTE: write this after rewriting the geospatial section. Also need to update conclusions.

3.3 Potential improvements

CAMELS-SPAT represents a substantial data processing effort, but further enhancements are possible. We briefly list these here. First, approximately 15% of our basin outlines have been assigned confidence ratings of medium or low. Future efforts can focus on refining these outlines, through further manual intervention, or higher resolution DEMs, or both. Second, we were somewhat limited in our ability to obtain hourly streamflow observations for the Canadian basins. Extension of these records would be helpful. Third, we necessarily needed to limit the extent of our geographical domain. However, all data sets used here have global coverage. Combination with local streamflow observations, and possibly high-quality local data sets, should allow for straightforward extension of the data set to other regions.

4 Conclusions

This paper describes the development of the CAMELS-SPAT data set. Our goal is to enable a wide range of hydrologic studies, with a particular focus on hydrologic modeling, by performing a wide range of data processing steps and sharing both the code and outcomes of these. We extend the original CAMELS data (Newman et al., 2015; Addor et al., 2017a) in three ways to achieve this goal. First, we extend the geographical domain of the data set beyond the contiguous United States by including Canadian basins. Second, we provided meteorological data specifically aimed at spatially-distributed physics-based hydrologic models, in addition to the inputs needed to run lumped, conceptual models. Third, we provide maps of multiple geospatial data sets for each basin, rather than a selection of summary statistics derived from these maps.

CAMELS-SPAT thus consists of meteorological data, streamflow observations and geospatial data for 1426 basins across the United States and Canada. The meteorological data includes a number of variables typically associated with process-based models, as well as potential evapotranspiration estimates that can be used with the more conceptual model types, at hourly time steps. This forcing data is provided in gridded format at its own resolution, as well as spatially averaged at the sub-basin and basin level. Streamflow observations are provided at daily time steps and complemented with hourly observations when these are available. Geospatial data, covering vegetation, land use, topography, soil and geology, are provided as geo-referenced maps for each basin, from which model inputs or summary statistics can easily be derived.

We made a conscious decision to limit the amount of subjectivity introduced in the data set by providing source data, rather than aggregated statistics. Instead, we focused on transparency, extensibility and efficiency of the data processing effort that CAMELS-SPAT represent. Our code is openly accessible under a permissive license. Coupled with the global extent of the forcing and geospatial data, creating preliminary CAMELS-SPAT products for other domains only involves processing of the relevant streamflow observations (though we readily acknowledge that including high-quality small-scale meteorological and geospatial data will enhance these potential data sets). By removing the need for a considerable amount of cumbersome data processing, we hope CAMELS-SPAT can support a wide range of hydrologic investigations at a fraction of the effort otherwise needed.

5 Code and data availability

Code needed to reproduce CAMELS-SPAT data preparation is available on GitHub with a fixed release available through Zenodo. Data source used in the preparation of this manuscript are listed in Table 4. Redistributed ERA5 data were generated using Copernicus Climate Change Service information [2023] in the case of the gridded forcing files. CAMELS-SPAT also contains modified Copernicus Climate Change Service information [2023] in the case of the (sub)basin-averaged forcing files. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains. We ask users to use the **citation.bib** file available on the CAMELS-SPAT data repository to cite each separate data set in any publications that use CAMELS-SPAT.

Appendix A: Timeseries resampling

Sub-daily streamflow observations provided by USGS and WSC are provided as instantaneous values (IV). Such values are only valid for the specific time at which they were obtained, and cannot be seen as representative for a time period on either side of the observation itself. This is illustrated in Figure A1, where a synthetic example of IV are shown as black circles and a simple approximation of the actual timeseries of streamflow as a black line. To accurately calculate the average flow over a particular time window, one first needs to integrate the IVs (under some assumption of how the flow changed between two consecutive IVs) to find the total volume of water that passed the gauge during the time window, and then divide this volume by the length of the time window to find the average flow rate for the time window. Such an approach is guaranteed to conserve

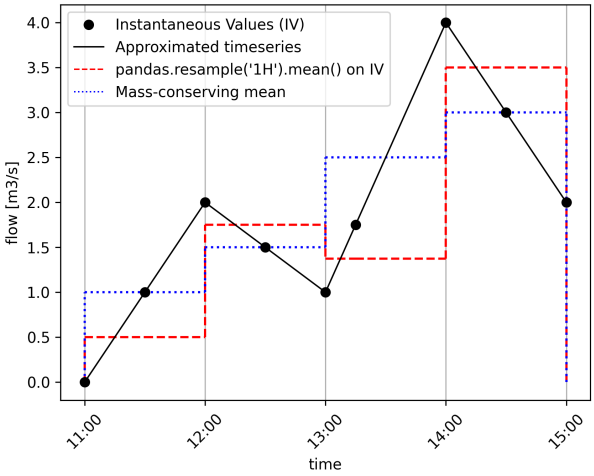

Figure A1. Example of standard resampling method applied to instantaneous values, as well as mass-conserving resampling.

.

mass (dotted blue line in Fig. A1). A potential pitfall is the direct application of tools such as the *resample('1H').mean()* method available in the popular Python package *pandas*. This method does not allow for "closed" windows (meaning that observations on both ends of the time window under consideration would contribute to the mean over the time window), and only provides options to include the "left" or "right" observation but not both. The method also does not take the length of the time window and where observations are located inside the time window into account (see the 13:00-14:00 window in Fig. A1). As a result, applying this method directly to the IV data leads to unintuitive average flow rate estimates, particularly when the IVs are not spaced equally in time (dashed red line in Fig. A1).

Appendix B: Streamflow data availability

Figure B1 shows streamflow data availability at a more granular level than the aggregated data in Figure 3.

Appendix C: Derivation of forcing variables

C1 Vapor pressure

ERA5 provides specific humidity and air pressure at the model level from which we download the data. We can derive vapor pressure as a function of specific humidity and air pressure as follows.

The ideal gas laws for dry air and water vapor are, respectively (Stull, 2017, Eq. 1.18 and 1.19):

$$P_d = \rho_d R_d T \tag{C1}$$

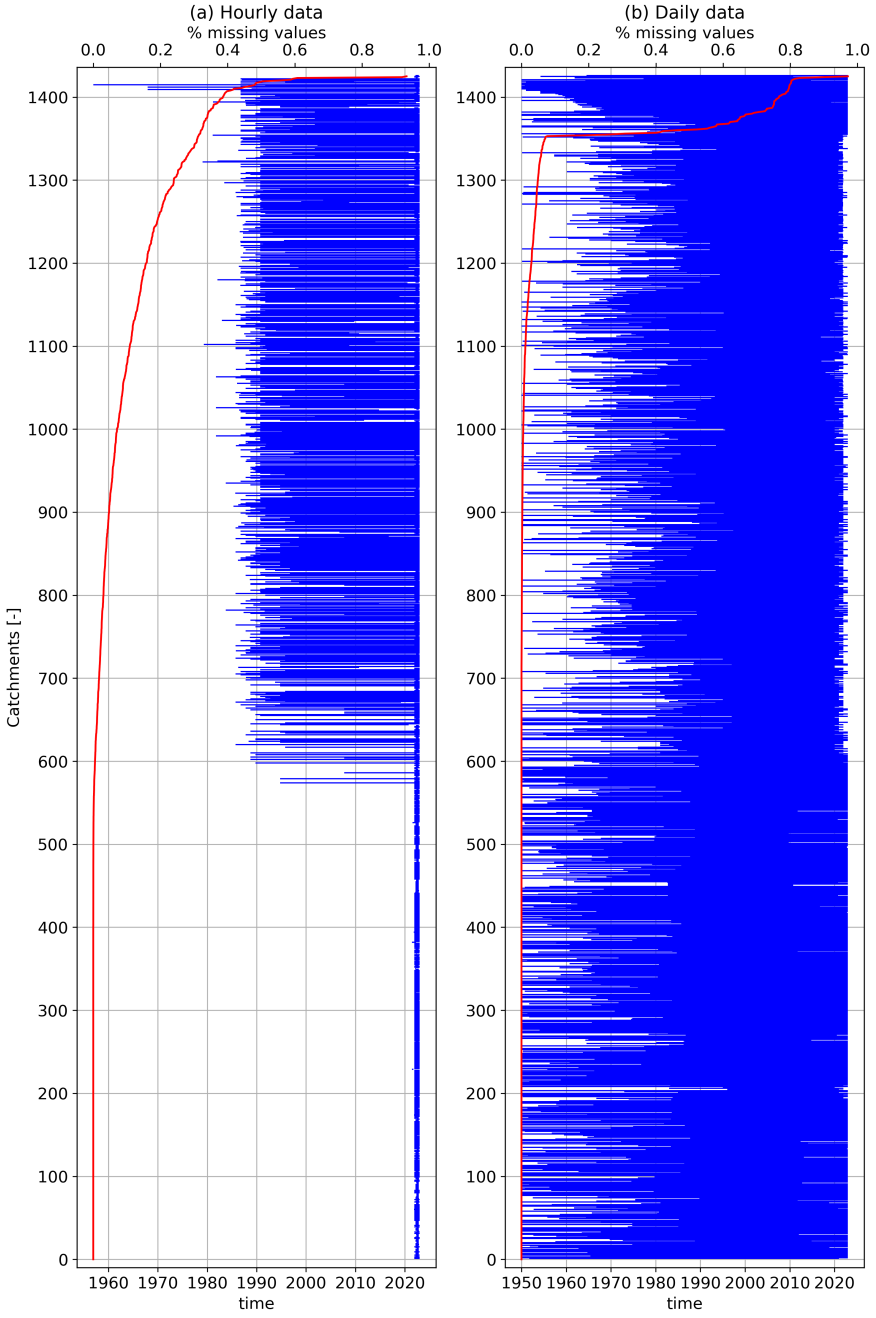

Figure B1. Flow data availability for gauges included in CAMELS-SPAT. The period on the lower x-axis refers to the period between the first publicly available flow record for a given station and its last, with this record period given in blue for each gauge. Missing values occur within this record period and are given here as percentages in red on the top x-axis.

$$e = \rho_v R_v T \tag{C2}$$

Where P_d and e are dry air pressure and vapor pressure $[kPa]$, respectively; ρ_d and ρ_v are the densities of dry air and water vapor $[kg\ m^{-3}]$; R_d and R_v are the gas constants for dry air and pure water vapor $[kPa\ K^{-1}\ m^3\ kg^{-1}]$; and T is temperature $[K]$.

Starting with the definition of specific humidity (Stull, 2017, Eq. 4.6):

$$q = \frac{m_v}{m_t} = \frac{m_v}{m_d + m_v} \tag{C3}$$

Where q is the specific humidity $[-]$, and m_v and m_t are the mass of water vapor and total mass of the air parcel $[kg]$, respectively. For the remainder, it makes sense to write the total air mass, m_t, as the sum of the mass of water vapor, m_v, and the mass of dry air, m_d, $[kg]$. Specific humidity can also be expressed as the ratio of densities, obtained by dividing both mass variables by their respective volumes V_v and V_d $[m^3]$:

$$q = \frac{\frac{m_v}{V_v}}{\frac{m_d}{V_d} + \frac{m_v}{V_v}} \tag{C4}$$

This is equivalent to the ratio of densities ρ $[kg\ m^{-3}]$:

$$q = \frac{\rho_v}{\rho_d + \rho_v} \tag{C5}$$

We can express the densities through their corresponding ideal gas laws:

$$q = \frac{\frac{e}{R_v T}}{\frac{P_d}{R_d T} + \frac{e}{R_v T}}, \tag{C6}$$

and reorganize these as follows:

$$q = \frac{e}{R_v T} \frac{1}{\frac{P_d}{R_d T} + \frac{e}{R_v T}} \tag{C7}$$

$$q = \frac{e}{R_v T} \frac{1}{\frac{P_d R_v T + e R_d T}{R_d R_v T^2}} \tag{C8}$$

$$q = \frac{e}{R_v T} \frac{1}{\frac{P_d R_v + e R_d}{R_d R_v T}} \tag{C9}$$

$$q = \frac{e}{R_v T} \frac{R_d R_v T}{P_d R_v + e R_d} \tag{C10}$$

$$q = \frac{R_v T}{R_v T} \frac{e R_d}{P_d R_v + e R_d} \tag{C11}$$

$$q = \frac{eR_d}{R_v\left(P_d + e\frac{R_d}{R_v}\right)} \tag{C12}$$

$$q = \frac{R_d}{R_v}\frac{e}{P_d + e\frac{R_d}{R_v}} \tag{C13}$$

$$\tag{C14}$$

The ratio of dry air and water vapor gas constants, $\frac{R_d}{R_v}$, is often given as ϵ:

$$q = \epsilon\frac{e}{P_d + e\epsilon} \tag{C15}$$

Assuming that total pressure $P = P_d + e$, we get an expression of specific humidity, q, in terms of total air pressure, P, vapor pressure, e, and gas constant ratio, ϵ (Stull, 2017, Eq. 4.7, top row):

$$q = \epsilon\frac{e}{P - e + e\epsilon} \tag{C16}$$

$$q = \frac{e\epsilon}{P - e\,(1-\epsilon)} \tag{C17}$$

In atmospheric sciences, this equation is often simplified to:

$$q \approx \frac{e\epsilon}{P} \tag{C18}$$

From which vapor pressure e can easily be obtained, under the assumption that the component $e(1-\epsilon)$ in the denominator is small compared to P. It is however possible to obtain an exact expression of e as a function of q, ϵ, and P. Starting from Eq. C17:

$$q\,(P - e\,(1-\epsilon)) = e\epsilon \tag{C19}$$

$$qP - qe + qe\epsilon = e\epsilon \tag{C20}$$

$$qe\epsilon - qe - e\epsilon = -qP \tag{C21}$$

$$e\,(q\epsilon - q - \epsilon) = -qP \tag{C22}$$

$$e = -\frac{qP}{q\epsilon - q - \epsilon} \tag{C23}$$

Where $\epsilon\,[-]$ has a constant value of 0.622, based on gas constants $R_d = 2.871{*}10^{-4}$ and $R_v = 4.61{*}10^{-4}\ [kPa\,K^{-1}\,m^3\,kg^{-1}]$.

C2 Relative humidity

ERA5 provides specific humidity and air temperature at the model level from which we download the data. We can use specific humidity to derive vapor pressure (see Appendix C1). With vapor pressure known, we can then derive relative humidity as follows (Stull, 2017, Eq. 4.14a):

$$RH = \frac{e}{e_s} \tag{C24}$$

Where RH is relative humidity $[-]$, e is vapor pressure $[kPa]$, and e_s is saturation vapor pressure at the current temperature $[kPa]$. e_s can be calculated using the Clausius-Clapeyron equation (Stull, 2017, Eq. 4.1a):

$$e_s \approx e_0 * exp\left[\frac{L}{R_v}\left(\frac{1}{T_0} - \frac{1}{T}\right)\right] \tag{C25}$$

Where e_s is saturation vapor pressure $[kPa]$, e_0 a known saturation vapor pressure at temperature T_0, L is a latent-heat parameter $[J\ kg^{-1}]$, R_v the water-vapor gas constant, and T air temperature $[K]$. Typical values are $e_0 = 0.6113\ [kPa]$ at $T_0 = 273.15\ [K]$. $R_v = 461\ [J\ K^{-1}\ kg^{-1}]$ (note the different units here compared to Appendix C1). The latent heat of vaporization for liquid water, $L_v = 2.5 \cdot 10^6\ [J\ kg^{-1}]$, while the latent heat of deposition for ice, $L_d = 2.83 \cdot 10^6\ [J\ kg^{-1}]$.

C3 Mean wind speed

ERA5 provides wind speed in the u and v directions. Mean wind speed w (in the same units as u and v) can be obtained with Pythagoras' theorem (European Centre for Medium-range Weather Forecasting, 2023a):

$$w = \sqrt{u^2 + v^2} \tag{C26}$$

C4 Wind direction

ERA5 provides wind speed in the u and v directions, from which a mean wind direction can be derived. We use here ECMWF's definition of *meteorological wind direction*, $\phi\ [degrees]$ which is the direction from which the wind blows, with North set at 0° and increasing clock-wise (European Centre for Medium-range Weather Forecasting, 2023a, see also the URL associated with this reference for a helpful graphic):

$$\phi = mod\left(180 + \frac{180}{\pi} atan2(v, u), 360\right) \tag{C27}$$

Author contributions. MC developed the idea for this data set and secured funding; NC provided guidance on geospatial data products; LRT provided assistance with geospatial data processing coding; WK developed the methodology, created the code, performed the data processing and wrote the initial draft of this paper; the paper was finalized with contributions of all co-authors.

Competing interests. TEXT

Disclaimer. TEXT

Acknowledgements. We are grateful to Louise Arnal for sharing her thoughts on early versions of some of our figures, to Chris Marsh for feedback on an early version of our methodology figure as well as pointing out some nuances about wind direction definitions, and to Guoqiang Tang for providing details about the way timestamps in the EM-Earth must be interpreted.

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, 2017a.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: Catchment attributes for large-sample studies, https://doi.org/10.5065/D6G73C3Q, 2017b.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrological Sciences Journal, 65, 712–725, https://doi.org/10.1080/02626667.2019.1683182, 2020.

Ahmed, M. I., Shook, K., Pietroniro, A., Stadnyk, T., Pomeroy, J. W., Pers, C., and Gustafsson, D.: Implementing a parsimonious variable contributing area algorithm for the prairie pothole region in the HYPE modelling framework, Environmental Modelling & Software, 167, 105 769, https://doi.org/10.1016/j.envsoft.2023.105769, 2023.

Almagro, A., Oliveira, P. T. S., Meira Neto, A. A., Roy, T., and Troch, P.: CABra: a novel large-sample dataset for Brazilian catchments, Hydrology and Earth System Sciences, 25, 3105–3135, https://doi.org/10.5194/hess-25-3105-2021, 2021.

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, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Arsenault, R., Bazile, R., Ouellet Dallaire, C., and Brissette, F.: CANOPEX: A Canadian hydrometeorological watershed database: CANOPEX: A Canadian Hydrometeorological Watershed Database, Hydrological Processes, 30, 2734–2736, https://doi.org/10.1002/hyp.10880, 2016.

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

Budhathoki, S., Rokaya, P., and Lindenschmidt, K.-E.: Improved modelling of a Prairie catchment using a progressive two-stage calibration strategy with in situ soil moisture and streamflow data, Hydrology Research, 51, 505–520, https://doi.org/10.2166/nh.2020.109, 2020.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth System Science Data, 12, 2075–2096, https://doi.org/10.5194/essd-12-2075-2020, 2020.

Clark, M. P. and Shook, K. R.: The Numerical Formulation of Simple Hysteretic Models to Simulate the Large-Scale Hydrological Impacts of Prairie Depressions, Water Resources Research, 58, e2022WR032 694, https://doi.org/10.1029/2022WR032694, 2022.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resources Research, 51, 2498–2514, https://doi.org/10.1002/2015WR017198, 2015a.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Gochis, D. J., Rasmussen, R. M., Tarboton, D. G., Mahat, V., Flerchinger, G. N., and Marks, D. G.: A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies, Water Resources Research, 51, 2515–2542, https://doi.org/10.1002/2015WR017200, 2015b.

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., Freer, J. E., Arnold, J. R., Moore, R. D., Istanbulluoglu, E., and Ceola, S.: Improving the theoretical underpinnings of process-based hydrologic models, Water Resources Research, 52, 2350–2365, https://doi.org/10.1002/2015WR017910, 2016.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, e2020WR029 001, https://doi.org/10.1029/2020WR029001, 2021.

Cloke, H. L. and Hannah, D. M.: Large-scale hydrology: advances in understanding processes, dynamics and models from beyond river basin to global scale, Hydrological Processes, 25, 991–995, https://doi.org/10.1002/hyp.8059, 2011.

Commission for Environmental Cooperation: North American Atlas – Political Boundaries, http://www.cec.org/north-american-environmental-atlas/political-boundaries-2021/, statistics Canada, United States Census Bureau, Instituto Nacional de Estadística y Geografía (INEGI), 2022.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020, 2020.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Horne, R., Whiteaker, T., Blodgett, D., Zender, C., Lee, D., Hassell, D., Snow, A. D., Kölling, T., Allured, D., Jelenak, A., Meier Soerensen, A., Gaultier, L., Herlédan, S., Manzano, F., Bärring, L., Barker, C., and Bartholomew, S.: NetCDF Climate and Forecast (CF) Metadata Conventions v1.11, http://cfconventions.org/Data/cf-conventions/cf-conventions-1.11/cf-conventions.html, accessed: 2024-01-11, 2023.

Environment and Climate Change Canada: National Water Data Archive: HYDAT, https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/national-archive-hydat.html, last Modified: 2018-07-05, 2010.

Environment and Climate Change Canada: Reference Hydrometric Basin Network, https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/reference-hydrometric-basin-network.html, last modified: 2021-02-26, 2020a.

Environment and Climate Change Canada: National hydrometric network basin polygons - Open Government Portal, https://open.canada.ca/data/en/dataset/0c121878-ac23-46f5-95df-eb9960753375, 2020b.

European Centre for Medium-range Weather Forecasting: ERA5: How to calculate wind speed and wind direction from u and v components of the wind? - Copernicus Knowledge Base - ECMWF Confluence Wiki, https://confluence.ecmwf.int/pages/viewpage.action?pageId=133262398, accessed: 2024-01-02, 2023a.

European Centre for Medium-range Weather Forecasting: ERA5: data documentation - Copernicus Knowledge Base - ECMWF Confluence Wiki, https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation#ERA5:datadocumentation-Howtoacknowledge,citeandrefertoERA5, accessed: 2024-01-03, 2023b.

European Centre for Medium-range Weather Forecasting: ERA5 terminology: analysis and forecast; time and steps; instantaneous and accumulated and mean rates and min/max parameters - Copernicus Knowledge Base - ECMWF Confluence Wiki, https://confluence.ecmwf.int/pages/viewpage.action?pageId=85402030#ERA5terminology:analysisandforecast;timeandsteps;instantaneousandaccumulatedandmeanratesandmin/maxparameters-Instantaneous,accumulated,meanrateandmin/maxparameters, accessed: 2024-01-03, 2023c.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology, 37, 4302–4315, https://doi.org/10.1002/joc.5086, 2017.

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., and Peel, M. C.: CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia, Earth System Science Data, 13, 3847–3867, https://doi.org/10.5194/essd-13-3847-2021, 2021.

Friedl, M. and Sulla-Menashe, D.: MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061, https://doi.org/10.5067/MODIS/MCD12Q1.061, 2022.

Gharari, S., Keshavarz, K., Knoben, W. J., Tang, G., and Clark, M. P.: EASYMORE: A Python package to streamline the remapping of variables for Earth System models, SoftwareX, 24, 101 547, https://doi.org/10.1016/j.softx.2023.101547, 2023a.

Gharari, S., Whitfield, P. H., Pietroniro, A., Freer, J., Liu, H., and Clark, M. P.: Exploring the provenance of information across Canadian hydrometric stations: Implications for discharge estimation and uncertainty quantification, preprint, Catchment hydrology/Instruments and observation techniques, https://doi.org/10.5194/hess-2023-150, 2023b.

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

Gleeson, T., Moosdorf, N., Hartmann, J., and Van Beek, L. P. H.: A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity, Geophysical Research Letters, 41, 3891–3898, https://doi.org/10.1002/2014GL059856, 2014.

Government of Canada: Web Service Links Interface - Water Level and Flow - Environment Canada, https://wateroffice.ec.gc.ca/services/links_e.html, see also the "guidelines" URL: https://collaboration.cmc.ec.gc.ca/cmc/hydrometrics/www/Document/WebService_Guidelines.pdf.

Government of Canada: Index of /cmc/hydrometrics/www/HydrometricNetworkBasinPolygons, https://collaboration.cmc.ec.gc.ca/cmc/hydrometrics/www/, note: accessed through https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/reference-hydrometric-basin-network.html, 2022.

Hamman, J. J., Nijssen, B., Bohn, T. J., Gergel, D. R., and Mao, Y.: The Variable Infiltration Capacity model version 5 (VIC-5): infrastructure improvements for new applications and reproducibility, Geoscientific Model Development, 11, 3481–3496, https://doi.org/10.5194/gmd-11-3481-2018, 2018.

Hansen, M. C., Potapov, P. V., Pickens, A. H., Tyukavina, A., Hernandez-Serna, A., Zalles, V., Turubanova, S., Kommareddy, I., Stehman, S. V., Song, X.-P., and Kommareddy, A.: Global land use extent and dispersion within natural land cover using Landsat data, Environmental Research Letters, 17, 034 050, https://doi.org/10.1088/1748-9326/ac46ec, 2022.

Hao, Z., Jin, J., Xia, R., Tian, S., Yang, W., Liu, Q., Zhu, M., Ma, T., Jing, C., and Zhang, Y.: CCAM: China Catchment Attributes and Meteorology dataset, Earth System Science Data, 13, 5591–5616, https://doi.org/10.5194/essd-13-5591-2021, 2021.

Hayashi, M., Van Der Kamp, G., and Rosenberry, D. O.: Hydrology of Prairie Wetlands: Understanding the Integrated Surface-Water and Groundwater Processes, Wetlands, 36, 237–254, https://doi.org/10.1007/s13157-016-0797-9, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: Complete ERA5 from 1940: Fifth generation of ECMWF atmospheric reanalyses of the global climate, https://doi.org/10.24381/cds.143582cf, 2017.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee,

D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., De Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, https://doi.org/10.24381/cds.adbb2d47, 2023.

Hogan, R.: Radiation Quantities in the ECMWF model and MARS, Tech. rep., European Centre for Medium-range Weather Forecasting, https://www.ecmwf.int/sites/default/files/elibrary/2015/18490-radiation-quantities-ecmwf-model-and-mars.pdf, 2015.

Höge, M., Kauzlaric, M., Siber, R., Schönenberger, U., Horton, P., Schwanbeck, J., Floriancic, M. G., Viviroli, D., Wilhelm, S., Sikorska-Senoner, A. E., Addor, N., Brunner, M., Pool, S., Zappa, M., and Fenicia, F.: CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland, preprint, ESSD – Land/Hydrology, https://doi.org/10.5194/essd-2023-127, 2023.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, 2005WR004 362, https://doi.org/10.1029/2005WR004362, 2006.

Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe, Earth System Science Data, 13, 4529–4565, https://doi.org/10.5194/essd-13-4529-2021, 2021.

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.2: an open- source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geoscientific Model Development, 12, 2463–2480, https://doi.org/10.5194/gmd-2018-332, 2019.

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, Scientific Data, 10, 61, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Lafaysse, M., Cluzet, B., Dumont, M., Lejeune, Y., Vionnet, V., and Morin, S.: A multiphysical ensemble system of numerical snow modelling, The Cryosphere, 11, 1173–1198, https://doi.org/10.5194/tc-11-1173-2017, 2017.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research, 99, 14 415, https://doi.org/10.1029/94JD00483, 1994.

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., David, C. H., Durand, M., Pavelsky, T. M., Allen, G. H., Gleason, C. J., and Wood, E. F.: Global Reconstruction of Naturalized River Flows at 2.94 Million Reaches, Water Resources Research, 55, 6499–6516, https://doi.org/10.1029/2019WR025287, 2019.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, Journal of Hydrology, 201, 272–288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Marsh, C. B., Pomeroy, J. W., and Wheater, H. S.: The Canadian Hydrological Model (CHM) v1.0: a multi-scale, multi-extent, variable-complexity hydrological model – design and overview, Geoscientific Model Development, 13, 225–247, https://doi.org/10.5194/gmd-13-225-2020, 2020.

Maurer, T. and Zehe, E.: CATFLOW: A Physically Based and Distributed Hydrological Model for Continuous Simulation of Catchment Water- and Solute Dynamics - User Guide and Program Documentation (Version CATSTAT), Tech. rep., INSTITUTE FOR WATER RESOURCES PLANNING, HYDRAULICS AND RURAL ENGINEERING (IWK), UNIVERSITY OF KARLSRUHE (TH), 2007.

Maxwell, R. M., Kollet, S. J., Condon, L. E., Smith, S. G., Woodward, C. S., Falgout, R. D., Ferguson, I. M., Engdahl, N., Hector, B., Lopez, S. R., Gilbert, J., Bearup, L., Jefferson, J., Collins, C., De Graaf, I., Prubilick, C., Baldwin, C., Bosl, W. J., Hornung, R., and Ashby, S.: PARFLOW User's Manual, Tech. rep., Integrated GroundWater Modeling Center, 2019.

McMillan, H., Coxon, G., Araki, R., Salwey, S., Kelleher, C., Zheng, Y., Knoben, W., Gnann, S., Seibert, J., and Bolotin, L.: When good signatures go bad: Applying hydrologic signatures in large sample studies, Hydrological Processes, 37, e14 987, https://doi.org/10.1002/hyp.14987, 2023.

Mekonnen, M. and Brauner, H.: MESH - A Community Hydrology-Land Surface Model: Meteorological Input, https://wiki.usask.ca/display/MESH/Meteorological+Input, accessed: 2022-01-27, 2020.

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O.: Estimating the volume and age of water stored in global lakes using a geo-statistical approach, Nature Communications, 7, 13 603, https://doi.org/10.1038/ncomms13603, 2016.

Mitchell, K., Ek, M., Wong, V., Lohmann, D., Koren, V., Schaake, J., Duan, Q., Gayno, G., Moore, B., Grunmann, P., Tarpley, D., Ramsay, B., Chen, F., Kim, J., Pan, H.-L., Lin, Y., Marshall, C., Mahrt, L., Meyers, T., and Ruscher, P.: THE COMMUNITY Noah LAND-SURFACE MODEL (LSM) - User's guide Public Release Version 2.7.1, Tech. rep., ftp://ftp.emc.ncep.noaa.gov/mmb/gcp/ldas/noahlsm/ver_2.7.1, 2005.

Muhammad, A., Evenson, G. R., Stadnyk, T. A., Boluwade, A., Jha, S. K., and Coulibaly, P.: Impact of model structure on the accuracy of hydrological modeling of a Canadian Prairie watershed, Journal of Hydrology: Regional Studies, 21, 40–56, https://doi.org/10.1016/j.ejrh.2018.11.005, 2019.

Myneni, R., Knyazikhin, Y., and Park, T.: MODIS/Terra+Aqua Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V061, https://doi.org/10.5067/MODIS/MCD15A2H.061, 2021.

National Weather Service: II.3-SAC-SMA: Conceptualization of the Sacramento Soil Moisture Accounting model, in: National Weather Service River Forecast System (NWSRFS) User Manual, pp. 1–13, http://www.nws.noaa.gov/ohd/hrl/nwsrfs/users_manual/htm/xrfsdocpdf.php,, 2005.

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R., and Blodgett, D.: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA, https://doi.org/10.5065/D6MW2F4D, artwork Size: approximately 2.5 GB Medium: text/plain, text/tab-separated-values, png, shp Pages: approximately 2.5 GB, 2014.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: 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.

Nijssen, B.: SUMMA Input - SUMMA Meteorological Forcing Files, https://summa.readthedocs.io/en/latest/input_output/SUMMA_input/#meteorological-forcing-files, 2017.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, Journal of Geophysical Research, 116, D12 109, https://doi.org/10.1029/2010JD015139, 2011.

Pellerin, J. and Nzokou Tanekou, F.: Reference Hydrometric Basin Network Update, Tech. rep., Environment and Climate Change Canada, Gatineau, QC, https://collaboration.cmc.ec.gc.ca/cmc/hydrometrics/www/RHBN/RHBN_EN.pdf, 2020.

Pelletier, J., Broxton, P., Hazenberg, P., Zeng, X., Troch, P., Niu, G., Williams, Z., Brunke, M., and Gochis, D.: Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers, p. 1032.940581 MB, https://doi.org/10.3334/ORNLDAAC/1304, artwork Size: 1032.940581 MB Publisher: ORNL Distributed Active Archive Center, 2016a.

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G., Williams, Z., Brunke, M. A., and Gochis, D.: A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling, Journal of Advances in Modeling Earth Systems, 8, 41–65, https://doi.org/10.1002/2015MS000526, 2016b.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

PIHM team: PennState Integrated Hydrologic Model (PIHM) - Version 2.0 - Input File Formats, Tech. rep., Hydrology Group, Civil & Environmental Engineering, Pennsylvania State University, http://www.pihm.psu.edu/Downloads/Doc/pihm2.0_input_file_format.pdf, 2007.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Pomeroy, J. W., Gray, D. M., Brown, T., Hedstrom, N. R., Quinton, W. L., Granger, R. J., and Carey, S. K.: The cold regions hydrological model: a platform for basing process representation and model structure on physical evidence, Hydrological Processes, 21, 2650–2667, https://doi.org/10.1002/hyp.6787, 2007.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C. E., Armston, J., Dubayah, R., Blair, J. B., and Hofton, M.: Mapping global forest canopy height through integration of GEDI and Landsat data, Remote Sensing of Environment, 253, 112 165, https://doi.org/10.1016/j.rse.2020.112165, 2021.

Rakovec, O., Kumar, R., Shrestha, P. K., and Samaniego, L.: Global assessment of hydrological components using a seamless multiscale modelling system, other, pico, https://doi.org/10.5194/egusphere-egu23-11945, 2023.

Schaake, J., Cong, S. Z., and Duan, Q. Y.: The US MOPEX data set, 307, 9–28, 2006.

Schulla, J.: Model Description WaSIM (Water balance Simulation Model), Tech. rep., Hydrology Software Consulting J. Schulla, http://www.wasim.ch/en/products/wasim_description.htm, 2021.

SMHI: HYPE file reference [HYPE Model Documentation], http://www.smhi.net/hype/wiki/doku.php?id=start:hype_file_reference#observation_data_files, accessed: 2022-01-27, 2022.

Stull, R. B.: Practical meteorology: an algebra-based survey of atmospheric science, Dept. of Earth, Ocean & Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada, version 1.02b edn., ISBN 978-0-88865-283-6, oCLC: 1054636700, 2017.

Tang, G., Clark, M., and Papalexiou, S. M.: EM-Earth: The Ensemble Meteorological Dataset for Planet Earth, https://doi.org/10.20383/102.0547, 2022a.

Tang, G., Clark, M. P., and Papalexiou, S. M.: EM-Earth: The Ensemble Meteorological Dataset for Planet Earth, Bulletin of the American Meteorological Society, 103, E996–E1018, https://doi.org/10.1175/BAMS-D-21-0106.1, 2022b.

Tarasova, L., Gnann, S., Yang, S., Hartmann, A., and Wagener, T.: Catchment characterization: current descriptors, knowledge gaps and future opportunities, preprint, Earth Sciences, https://doi.org/10.31223/X5BM2G, 2023.

Teluguntla, P., Thenkabail, P., Oliphant, A., Gumma, M., Aneece, I., Foley, D., and McCormick, R.: Landsat-Derived Global Rainfed and Irrigated-Cropland Product 30 m V001, https://doi.org/10.5067/COMMUNITY/LGRIP/LGRIP30.001, 2023.

Thenkabail, P. S., Teluguntla, P. G., Xiong, J., Oliphant, A., Congalton, R. G., Ozdogan, M., Gumma, M. K., Tilton, J. C., Giri, C., Milesi, C., Phalke, A., Massey, R., Yadav, K., Sankey, T., Zhong, Y., Aneece, I., and Foley, D.: Global Cropland-Extent Product at 30-m Resolution

(GCEP30) Derived from Landsat Satellite Time-Series Data for the Year 2015 Using Multiple Machine-Learning Algorithms on Google Earth Engine Cloud, USGS Numbered Series, U.S. Geological Survey, https://doi.org/10.3133/pp1868, series: Professional Paper, 2021.

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., and Zhang, Y.: Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States, Hydrology and Earth System Sciences, 27, 1809–1825, https://doi.org/10.5194/hess-27-1809-2023, 2023.

Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., and Peel, M. C.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.1: an object-oriented implementation of 47 established hydrological models for improved speed and readability, Geoscientific Model Development, 15, 6359–6369, https://doi.org/10.5194/gmd-15-6359-2022, 2022.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset, Water Resources Research, 55, 5053–5073, https://doi.org/10.1029/2019WR024873, 2019.

Table 4. Data sources, references, licenses and links for further information.

Type	Name	Reference(s)	License
Flow direction grid	Merit Hydro Adjusted Elevations	Yamazaki et al. (2019)	CC-BY-NC 4.0 or ODbL 1.0
Flow accumulation grid	Merit Hydro Adjusted Elevations	Yamazaki et al. (2019)	CC-BY-NC 4.0 or ODbL 1.0
Sub-basin polygons	MERIT Basins	Lin et al. (2019)	[1]
Meteorological forcing	ERA5	Hersbach et al. (2020, 2017, 2023)	Copernicus Data License[3]
Meteorological forcing	Deterministic EM-Earth	Tang et al. (2022b, a)	CC-BY 4.0
Climate	WorldClim	Fick and Hijmans (2017)	Derived data only, under CC-BY
Forest height	Global Land Cover and Land Use Change, 2000-2020	Potapov et al. (2021)	CC-BY
Leaf Area Index	MCD15A2H.061	Myneni et al. (2021)	No restrictions[4]
Agriculture	LGRIP30	Thenkabail et al. (2021); Teluguntla et al. (2023)	No restrictions[4]
Land cover & land use	MCD12Q1.061	Friedl and Sulla-Menashe (2022)	No restrictions[4]
Land cover & land use	Global land cover and land use 2019	Hansen et al. (2022)	CC-BY 4.0
Lakes	HydroLAKES	Messager et al. (2016)	CC-BY 4.0
Digital Elevation Model	Merit Hydro Adjusted Elevations	Yamazaki et al. (2019)	CC-BY-NC 4.0 or ODbL 1.0
Soil properties	SOILGRIDS 2.0	Poggio et al. (2021)	CC-BY-NC 4.0
Soil properties	Pelletier	Pelletier et al. (2016b, a)	No restrictions[5]
Geology	GLHYMPS	Gleeson et al. (2014); Gleeson (2018)	CC-BY 4.0

[1] No formal license stated in the paper, but data is publicly available for research purposes

[2] Original link no longer accessible; new data location unknown

[3] See: https://cds.climate.copernicus.eu/api/v2/terms/static/licence-to-use-copernicus-products.pdf (accessed: 2023-12-18)

[4] See: https://lpdaac.usgs.gov/data/data-citation-and-policies/ (accessed: 2023-10-17)

[5] See: https://www.earthdata.nasa.gov/learn/use-data/data-use-policy (accessed: 2023-12-18)