https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

Open Access Earth System Science Data Discussions

GRDC-Caravan: extending Caravan with data from the Global Runoff Data Centre

Claudia Färber[1], Henning Plessow[1], Simon Mischel[1], Frederik Kratzert[2], Nans Addor[3,4], Guy Shalev[5], Ulrich Looser[1]

[1] Global Runoff Data Centre (GRDC), Federal Institute of Hydrology (BfG), Koblenz, 56068, Germany
[2] Google Research, Vienna, 1010, Austria
[3] Fathom, Bristol, BS8 1EJ, UK
[4] University of Exeter, Exeter, EX4 4RJ, UK
[5] Google Research, Tel Aviv, 6789141, Israel

Correspondence to: Claudia Färber (faerber@bafg.de)

Abstract. Large-sample datasets are essential in hydrological science to support modelling studies and advance process understanding. However, these datasets often lack standardization, which impedes their combination. Caravan is a community initiative to create a large-sample hydrology dataset of meteorological forcing data, catchment attributes, and discharge data for catchments around the world. Compared to existing large-sample hydrology datasets, the focus of Caravan is to use globally consistent forcing and attribute data to facilitate global studies. Caravan is a community project designed to be expanded by members of the hydrological community using a common cloud-based framework. This dataset is currently the 6th extension to Caravan, based on a subset of hydrological discharge data and station-based watersheds from the Global Runoff Data Centre (GRDC), which are covered by an open data policy. The GRDC is an international data centre operating under the auspices of the World Meteorological Organization (WMO), which collects quality-controlled river discharge data and associated metadata from the National Meteorological and Hydrological Services (NMHS) of WMO Member States. The dataset covers stations from 5,356 catchments and 25 countries, spans the years 1950 – 2023. This takes the total number of Caravan catchments (core dataset plus extensions) to 22,372 (1589 catchments accounting for duplicates). This extension is released under a CC-BY-4.0 license that allows redistribution and is publicly available on Zenodo: https://zenodo.org/records/14006282 (Färber et al., 2024). We encourage additional NMHS to make their data available under open licenses so that it can be included in future versions of the extension.

1 Introduction

River systems are an integral part of the global water cycle, which are linked to many processes on local, regional and global scales (Dorigo et al., 2021). Observational river discharge data are counted as Essential Climate and Water Variables being fundamental for a wide range of applications such as flood and drought management, the modelling of the global water balance, the analysis of long-term circulation patterns, and the estimation of fluxes into the oceans (Lawford et al., 2023; GCOS, 2021;

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

GEO, 2014). The collection and dissemination of long-term hydrological data and information on a global scale is the primary objective ~~of the principal reason for the operation of~~ the Global Runoff Data Centre (GRDC). The GRDC is an international data centre, which operates under the auspices of the World Meteorological Organization (WMO) at the German Federal Institute of Hydrology (BfG). Established in 1988 to support the research on global and climate change and integrated water resources management, GRDC holds the most substantive collection of quality assured river discharge data on a global scale. Primary providers of the river discharge data are the NMHS of the countries, coordinated by WMO. Currently, the Global Runoff Database contains river discharge data collected at daily or monthly intervals from more than 10,000 stations in 160 countries, dating back up to 200 years. GRDC provides river discharge data for non-commercial use, but all data remain property of the owner and redistribution is not allowed (GRDC, 2024a).

In parallel to hydrological data made available by GRDC and water services (e.g., US Geological Survey, USGS), river discharge data is increasingly being provided through large-sample hydrology (LSH) datasets collated by third parties. There are global collections of streamflow data such as the Global Streamflow Indices and Metadata Archive (GSIM), which provides global streamflow indices and metadata (Do et al., 2018; Gudmundsson et al., 2018). Other datasets, such as MOPEX (Schaake et al., 2006), CAMELS (Newman et al., 2015; Addor et al., 2017) and HYSETS (Arsenault et al., 2020) are based on flow time series, which are complemented by atmospheric forcing (e.g., precipitation) time series and catchment attributes (characterizing e.g., the dominant soil type or the level of anthropogenic water extractions). These time series and attributes opened new possibilities in hydrology, enabling for instance the development of machine learning based hydrological models (e.g. Kratzert et al., 2019a; Kratzert et al., 2019b) and the identification of human impacts on river flows over large regions (e.g. Bloomfield et al., 2021; Chagas et al., 2022).

While these datasets have led to key advances, progress is hindered by a few shortcomings (Addor et al., 2020). One of them is the lack of common standards, which makes their combination challenging. The Caravan project (Kratzert et al., 2023) was designed to overcome this and to create a global dataset by combining seven existing LSH datasets (Newman et al., 2015; Fowler et al., 2021; Chagas et al., 2020; Alvarez-Garreton et al., 2018; Coxon et al., 2020; Arsenault et al., 2020; Klingler et al., 2021). Caravan is a truly global open-source community resource, which allows members of the hydrology community to extend the dataset to new locations through a cloud-based workflow (Fig. 1). Meteorological forcing data and catchment attributes for new Caravan catchments are thereby extracted from the global datasets ERA5-Land (Muñoz-Sabater et al., 2021) and HYDRO-Atlas (Lehner et al., 2019; Linke et al., 2019), guaranteeing spatial consistency and comparability across regions. Since its release, several extensions have been produced covering different spatial and hydroclimatic regions of the world (Koch, 2022; Morin, 2024; Höge et al., 2023; Casado Rodríguez, 2023; Helgason and Nijssen, 2024).

Here we use the Caravan framework to generate a new extension based on river discharge time series from the Global Runoff Database. A key criterion to select catchments included in this extension was that their river discharge data could be made

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

Open Access Earth System Science Data Discussions

publicly available. Because the license applying to the whole Global Runoff Database is restrictive (i.e. data can only be used for non-commercial purposes and cannot be redistributed), here only gauge data with a permissive license were selected. Hence, this extension could be released under CC-BY-4.0, like the majority of Caravan. In other words, the license reflects that set by the data owner (the NHMS) and data is not redistributed without their consent.

2 Dataset description

The GRDC-Caravan extension was developed following the core Caravan methodology in Kratzert et al. (2023). It contains meteorological forcing data, catchment attributes and river discharge data (Fig. 1). Meteorological forcing data and catchment attributes are derived from ERA-5 Land times series (Muñoz-Sabater et al., 2021) and HydroATLAS (Lehner et al., 2019; Lehner et al., 2022; Linke et al., 2019), respectively. River discharge data and catchment polygons are compiled by GRDC. The extraction of attributes and time series has been carried out using Google Earth Engine in the cloud. A detailed description of the different components is provided below.

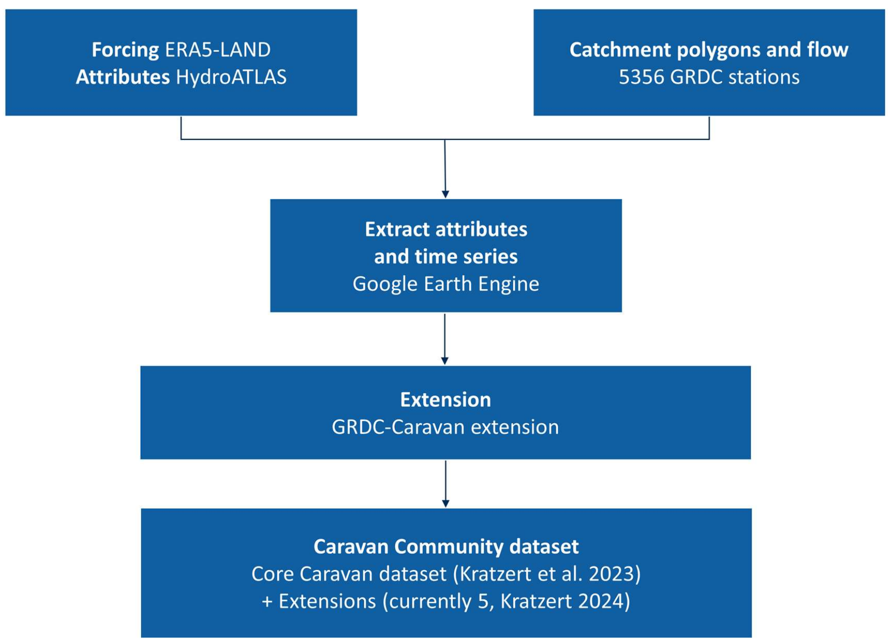


Figure 1: Flow-chart demonstrating the Caravan framework introduced in Kratzert et al. (2023). See Kratzert (2024) for published Caravan extensions.

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

2.1 River discharge data

The extension covers stations from 5356 catchments and 25 countries from the Global Runoff Database, spanning a time frame from 1950 – 2023 (Fig. 2, Table 1). Following Kratzert et al. (2023) and GRDC's data policy, only those stations whose countries adhere to an open data policy with a permissive data sharing license were included in the dataset. In addition, only stations supported by a polygon of the catchment area were chosen. Furthermore, the stations must contain daily discharge values with a time series of at least 1 year and should end in 2010 or more recently. As for the rest of Caravan, discharge data is provided in area normalized units (mm/day). We used the area of the catchment polygons for the normalization, which is provided as the area in the catchment attribute.

Similar to the base Caravan dataset, most catchments of the extension are located in WMO region IV (North America, Central America and the Caribbean) and WMO region VI (Europe) (Fig. 2, Table 1). While in the base dataset distinctly more catchments are available in the United States, the GRDC Caravans extension provides several new stations in Canada and Europe. In WMO region III (South America), more catchments are available in the base Caravan dataset, mostly due to data from Chile, but the GRDC extension provides additional data from Brazil and Argentina. A large increase in stations could also be reached in WMO region I (Africa) and Region V (South-West Pacific). While in the base dataset there are no stations in Africa and only a few in Australia, the extension provides significantly more stations from Australia, South Africa, New Zealand, Namibia and Liberia. Solely from WMO Region II (Asia) no data could be provided through the extension and there are also no stations available from this region in the original Caravans dataset. With respect to climate zones, most catchments in Caravan are still available from cool and cold temperate regions (Fig. 3). But with data from South Africa and Australia, a distinct increase in catchments from extremely hot regions could be reached through the extension.

Figure 4 and 5 provide an overview about the distribution of streamflow records through time and their average lengths. The highest number of gauging stations in the extension is available between 1980-2010, with about 1000 stations providing more than 70 years of data. The average length of records is about 50 years (Table 1). The dataset shows a high data completeness with only few missing values (Table 1), whereby parts of the annual fluctuation in the data availability can be attributed to missing data in snow-dominated regions during winter (Fig. 4). In the same figure, the strong decline of streamflow records since 2020, however, must not be misinterpreted. It is not an indication of a decreasing monitoring infrastructure, but is due to the fact that GRDC only includes quality controlled (yearbook) river discharge data from NHMS. Since many countries take several years for the final release of the quality-controlled data, this time lag until GRDC can release quality-controlled data with confidence is evident in the drop-off of the number of available stations released by NMHS (Fig. 4).

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

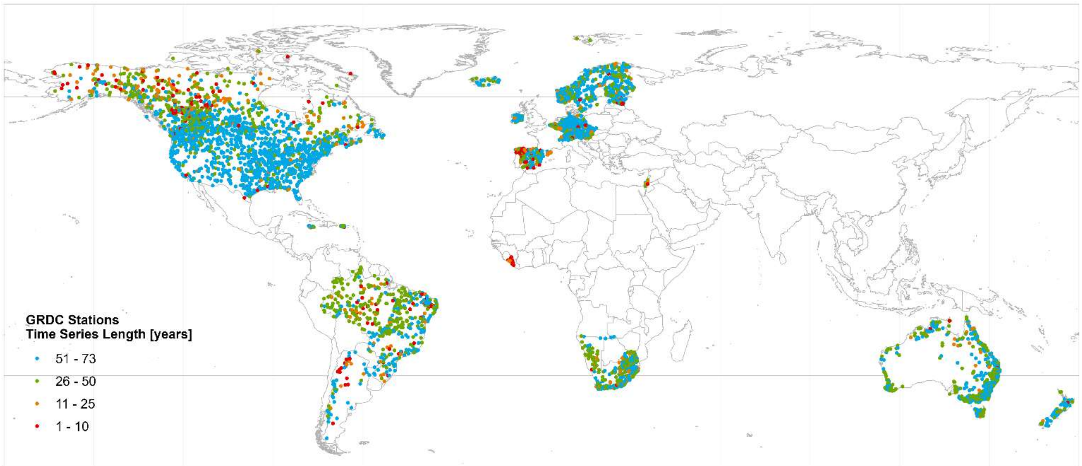


Figure 2: Global distribution of catchments included in the GRDC-Caravan extension indicated by river discharge time series length. Note that the data records in the Caravan extension starts 1951 (because of the forcing data). Longer records for individual stations might exist on the GRDC data portal (GRDC, 2024b).

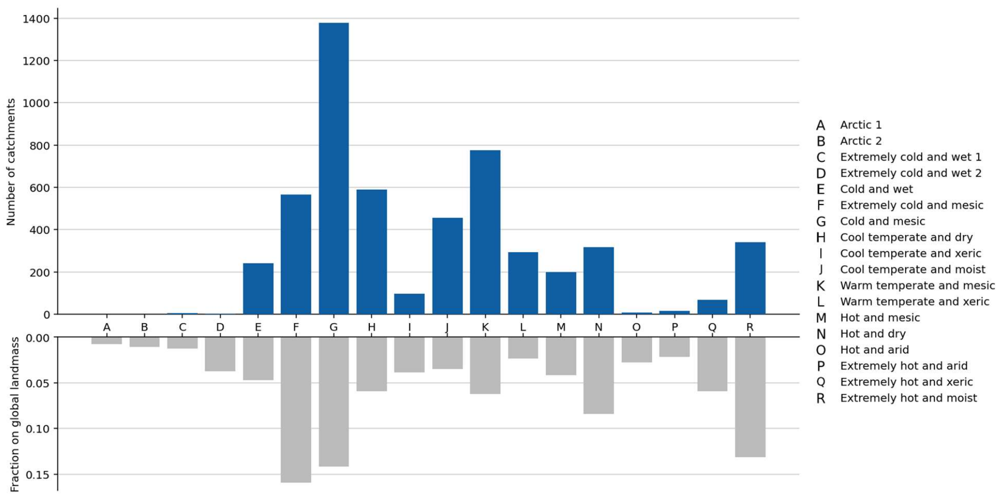
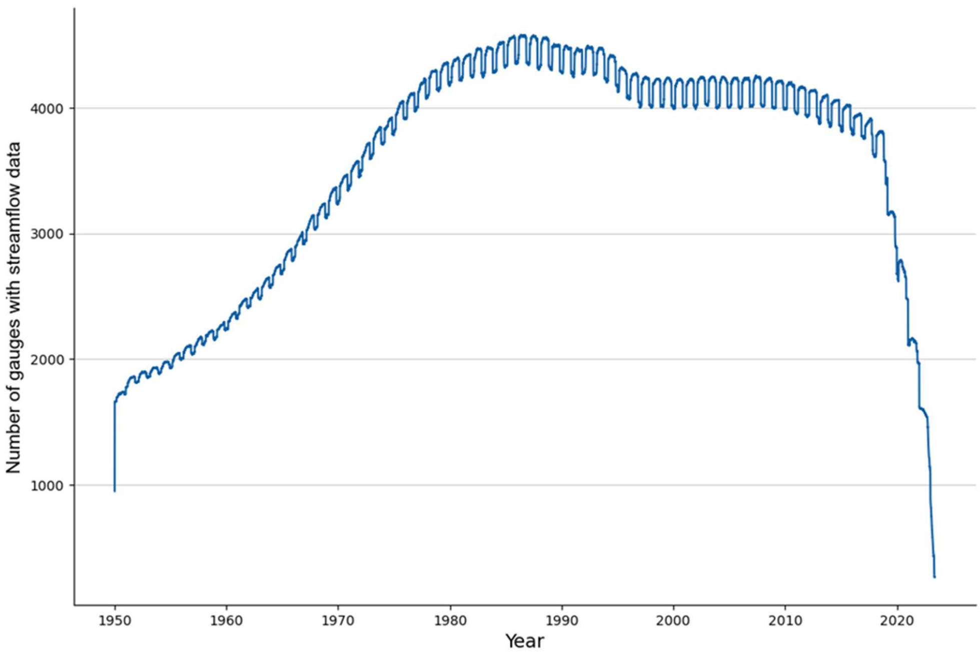


Figure 3: Distribution of catchments among the Global Environmental Stratification (GEnS) climate zones (Metzger et al., 2013). The bottom part of the plots shows the fraction of a particular climate zone on the total land mass.

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024



120 **Figure 4: Number of gauging stations with streamflow data through time. The recent drop-off is not an indication of a decreasing monitoring infrastructure, but rather is evidence of the time it takes until GRDC receives data from NMHS and can release quality-controlled data. The annual cycles in the data availability are due to missing data in snow-dominated regions during winter.**

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

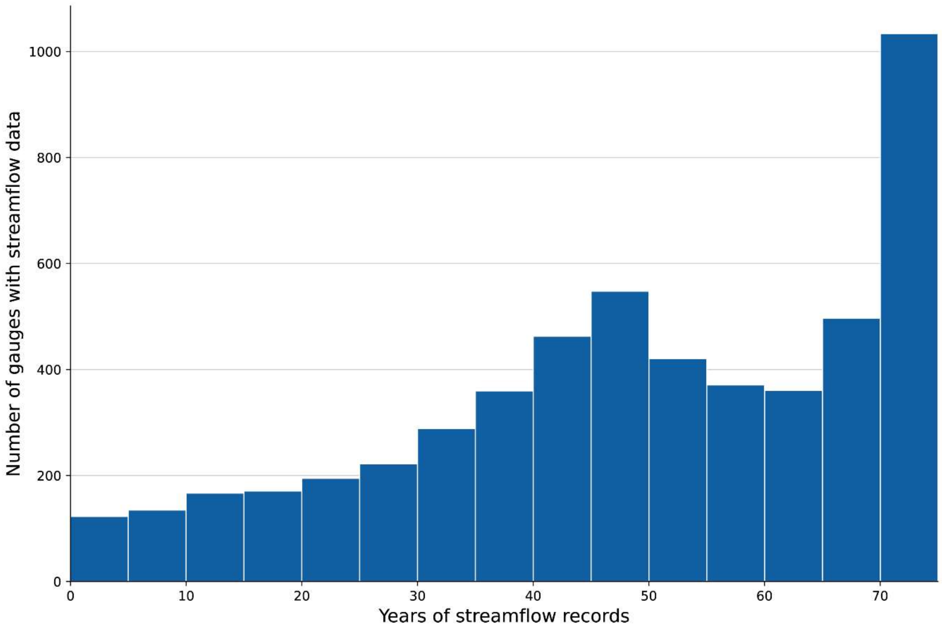


Figure 5: Number of gauging stations with streamflow record over time (average times series length of about 50 years).

Table 1: Overview of country data included into the dataset. From WMO Region II (Asia) no data is provided in the extension.

Region/Country	**Number of stations**	**Average times series length [yrs]**	**Data completeness [%]**
WMO Region I: Africa			
Liberia	9	8.5	98.3
Namibia	51	45.2	92.8
South Africa	434	48.4	93.4
WMO Region III: South America			
Argentina	59	43.0	94.7
Brazil	472	41.7	93.1

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

WMO Region IV: North America, Central America, Caribbean			
Canada	1133	44.7	85.4
Jamaica	12	51.9	82.2
Puerto Rico	24	45.4	95.5
United States	999	63.3	95.6
WMO Region V: South-West Pacific			
Australia	701	49.8	97.0
New Zealand	66	50.4	98.4
WMO Region VI: Europe			
Austria	134	54.6	99.7
Belgium	55	34.0	96.8
Czech Republic	29	68.1	99.6
Estonia	51	49.5	89.4
Finland	133	54.3	99.1
Germany	336	64.1	99.8
Iceland	23	55	95.9
Ireland	43	54.6	92.9
Israel	8	33.3	97.2
Luxembourg	11	18.0	100
Norway	206	51.8	96.4
Spain	145	41.8	91.5
Sweden	130	57.8	98.6
Switzerland	92	56.2	99.8

2.2 Catchment polygons and gauge locations

All stations that meet the criteria for inclusion in the Caravan dataset, contain a polygon of the station's catchment area. The polygons are mandatory and play a crucial role in the Caravan dataset, as they are used to derive meteorological forcing data (see chapter 2.3). Below a detailed description of the calculation and evaluation of the catchment areas and the station allocation is given. Additionally, a quality indicator for the calculated catchments is provided.

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

Open Access Earth System Science Data Discussions

2.2.1 Area calculation

For the calculation of station catchment areas, a global flow direction raster was chosen, preferably with an associated river network. The initial dataset utilized is the HydroSHEDS dataset with a 15 arc-second resolution, which corresponds to approximately 500 m at the equator. The dataset has almost global coverage, but was supplemented with the Hydro1k data set above 60° northern latitude, where the quality for these catchment areas drops considerably (Lehner et al., 2008). Therefore, the Multi-Error-Removed Improved-Terrain (MERIT) Hydro dataset was used additionally. MERIT has a 3 arc-second resolution, equivalent to about 90 meters at the equator (Yamazaki et al., 2019). This dataset is well-suited for higher latitudes and particularly for smaller catchment areas. Lin et al. (2019) employed the TauDEM tool to derive the river network of MERIT Hydro.

For the calculation of the upstream area of the stations with the HydroSHEDS dataset the self-developed R package ‘GRDCFlowTools’ has been used. It contains functions to process raster files and to determine all upstream cells of a given point, using among others the R package *igraph* (Csardi and Nepusz, 2006), which is a library for graph theory and network analysis based on the principle of topological sort. Creating these graph objects requires a lot of computing memory and quickly reaches its limit, when using raster sets covering a whole continent. The function *createSplitFlowGraph* of the self-developed R package was used, to divide a flow direction raster into any number of stripes including their graph objects. For these calculations a separation of the HydroSHEDS flow direction raster into 10 parts seemed to be a reasonable amount. With the created graph-objects the function *getBasinFromSplitGraph* derives all inflow cells to a given cell number of a flow direction raster from which the total area was determined. This procedure was done for each pixel on the river network.

For the area calculations with the MERIT Hydro dataset the *delineator.py* Python scripts of Heberger (2021) were used. It allows quick watershed delineation, using a hybrid of vector- and raster-based methods. It consists of two modes: A “high-resolution” mode, based on MERIT Hydro and a “low-resolution” mode based on HydroSHEDS. We used, the “high-resolution” mode, which requires only a column-based file with a unique id and the coordinates to run the script. The huge advantage of the method is how fast the raster is read and processed. This is due to the python package ‘*pysheds*’, which is able to read the raster files using a bounding box, where only a particular part of the raster grid is processed. The bounding box is thereby created by the intersection of the unit catchment and the outlet point, clipping the flow direction and accumulation grids to the extent of the unit catchment. To achieve this, all cells outside the unit catchment are marked as 0 to prevent that a neighbouring catchment is considered. Before the catchment is calculated with the *grid.catchment* function of *pysheds* the point is snapped to the nearest stream, which depends on the number of upstream pixels to define a waterway. For smaller catchments a smaller number is recommended and vice versa for large ones.

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

Open Access Earth System Science Data Discussions

2.2.2 Catchment evaluation and station allocation

The methodology for calculating and evaluating the station catchment areas was adopted from the technical report of Lehner (2012). To link the gauging stations to the river network, a two-fold strategy was used. First an automatic station allocation was carried out, where a 5 km buffer around each station has been drawn. Within this buffer, points for each cell of the flow direction raster have been created. Those cells were then clipped to the river network, if existent, which reduces the number of points considerably. After the calculation of all potential pour points was completed, a rating score R was calculated with the following equation:

$$\mathbf{R = RA + 2RD} \tag{1}$$

, where RA is the area discrepancy and RD the distance ranking. RA was calculated by a relative comparison between the calculated watershed area and the original GRDC catchment area provided by the NMHS, with a value of 0 indicating a perfect match. A positive or negative deviation until 50 % was tolerated and assigned with the value 50, while all stations that exceeded this tolerance limit were initially sorted out and inspected later in detail manually. RD was derived from the distance of the station's coordinates to the respective pour point. Due to the double weighted distance ranking, stations are less likely to be chosen if they are far away. A distance of 2 km would have a value of 40. Since the buffer was 5 km large, a maximum value of 100 could be obtained. The two best results, namely those with the best overall ranking calculated from MERIT and HydroSHEDS were chosen, and the catchment polygon was calculated based on the one with the lowest R-value (Eq. 1). These stations were assigned as type “automatic”, while all other stations that exceeded the ranking of 150 were inspected manually and assigned as type “manual”. Manual procedures for station allocation involved the verification of river and station names as well as coordinates. The final decision on whether a station was moved to a new and “reliable” location was also based on the agreement of the modelled and reported GRDC catchment area and the long-term annual discharge.

Table 2: Differentiation of quality levels of calculated catchments.

Type	Quality	Comment	Number of stations
Automatic	High	Area discrepancy <= 5 % and distance <= 5 km	3918
	Medium	Area discrepancy 5-10 % and distance <= 5 km	613
	Low	Area discrepancy 10-50 % and distance <= 5 km	691
Manual	High	Station and river name could be identified, area discrepancy <= 5 %	55
	Medium	Station or river name could be identified, area discrepancy between 5-10 %	9
		Location seems correct, but GRDC area seems wrong	18
	Low	Stations were relocated manually, catchment agreement between 10-50 %, mediocre correlation with the flow behaviour	46
		Location ok, but catchment not well represented	6

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

2.2.3 Catchment quality indicator

Even if a minimum quality is ensured with the help of the ranking assessment, the question for the user is how reliable the derived station catchment areas are and how good the agreement is, since not all station catchment areas have the same quality. As the quality of the catchment boundaries depend on the accuracy of the provided coordinates as well as the catchment area, a further quality indicator has been assigned. The differentiation was based on whether the stations were generated automatically or manually and was further divided into three quality levels: “High”, “Medium” and “Low”. (Table 2). The quality indicator of catchments, which were calculated automatically was assigned according to the calculated area discrepancy and distance. “High” quality was assigned for catchments with an area discrepancy <= 5 % and a distance <= 5 km, “Medium” quality for an area discrepancy 5-10 % and a distance <= 5 km, and “Low” quality for an area discrepancy between 10-50 % and a distance <= 5 km.

The quality of manually calculated catchments was more complicated. Manually calculated catchments were assigned with a "High" quality index, if both the station and river name could be identified and the area discrepancy is less or equal 5 %. If either the station or river name could be identified but the area discrepancy is between 5-10 %, catchments were classified into the “Medium” quality category. The “Medium” category was also assigned for cases, where the location seems correct, but it is evident that the provided area is incorrect, since regional case studies or comparisons with upstream or downstream stations reveal discrepancies. “Low” quality involved catchments, were some of the stations were relocated manually and the catchment agreement tends to be mediocre between 10-50 %, but correlates to some extent with the flow behaviour. The derivation was not always clear, especially for stations located in arid regions. In addition, catchments were assigned to the “Low” quality category, whose coordinates are correct, but the elevation model could not represent the catchment accurately.

In total the catchment area for 5,356 stations were calculated, with 5,222 stations being allocated automatically and 134 manually inspected stations (Table 2). The distribution of the catchment areas is shown in Fig. 6. The majority of the stations (85 %) have a catchment area size of 1,000-10,000 km² followed by stations with catchment area sizes of 100-1,000 km² and 10,000-100,000 km². Only few catchment areas are smaller than 10 or 100 km², or larger than 100,000 km².

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

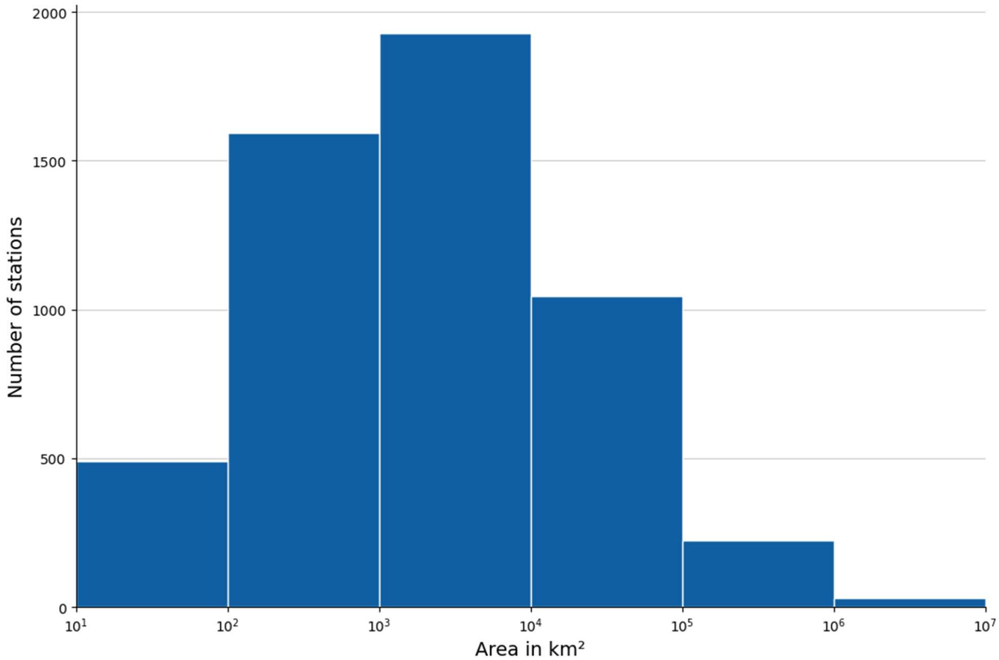


Figure 6: Histogram of area distribution.

215 ### 2.3 Meteorological forcing data

The GRDC-Caravan dataset contains the same 38 ERA5-Land time series features derived with the same algorithm as the original Caravan dataset, see Table 1 and Table 2 in Kratzert et al. (2023). Note that the potential evaporation band from ERA5-Land, which is included in Caravan and this extension, is known to have issues, see e.g. Clerc-Schwarzenbach et al. (2024). For consistency, this band is still included in our extension, but it should be used with care.

220 ### 2.4 Catchment attributes

The GRDC-Caravan extension includes the same 196 catchment attributes derived from HydroATLAS (see Table 3 and Table 4 in Kratzert et al. (2023), the same 10 climate indices derived from the meteorological time series (see Table 5 in Kratzert et al. (2023), as well as the same metadata information (gauge latitude, longitude, area, country, and station name; see Table 6 in

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

Kratzert et al. (2023)). For better comparability, the climate indices were calculated from the same time period (1981-2022) as in the original Caravan dataset, even though longer records of forcing data are available.

Since the GRDC-Caravan extension also includes catchments that are much larger than the upper threshold that was used in the original Caravan dataset, which was 2000 km^2, we had to adapt the code that is responsible for deriving the catchment attributes from HydroATLAS on Google Earth Engine. The updated code has been merged into the official Caravan GitHub repository (Kratzert, 2024).

Additionally, the GRDC-Caravan extension includes a set of additional attributes that are specific to the GRDC data (see Table 3). Most of the attributes are derived from the GRDC station catalogue, available at the GRDC data portal (GRDC, 2024b). However, we made sure that the attributes related to streamflow availability are clipped to the periods included in this extension and that long-term streamflow indices are derived equally from the data that is included in the extension. Another attribute that was added is *nat_id*, which is the national station code for each gauge station. This allows the user to find duplicates between GRDC station codes and stations that are already included in Caravan.

The reason is that we also provide a mapping from the GRDC gauge ID to the national station IDs, which helps to filter out duplicates between GRDC stations and stations from other datasets already included in Caravan. In total there are 1589 duplicates. 202 with CAMELS-BR, 944 with HYSETS, 69 with CAMELS (US), 191 with CAMELS-AUS, and 183 with LamaH-CE. There are a few gauges where a single GRDC gauge maps to both, a CAMELS (US) and a HYSETS gauge, which are counted as two duplicates in the value 1589. The high number of duplicates were still included in the expansion because they add an average of 13.8 time series years.

Table 3: Additional attributes that are included in the GRDC-Caravan extension.

Attribute	Description	Unit
grdc_no	GRDC station number	-
nat_id	National station id	-
wmo_reg	WMO region	-
sub_reg	WMO subregion (basin)	-
country	Country code (ISO 3166)	-
area_shp	Catchment size, as derived catchment polygons	km^2
altitude	Height of gauge zero	m.asl
d_start	Daily data available since	year
d_end	Daily data available until	year
d_yrs	Length of time series of daily data	-
d_miss	Percentage of missing values	-

https://doi.org/10.5194/essd-2024-427
Preprint. Discussion started: 13 November 2024

lta_discharge	Long-term average discharge	m^3s^{-1}
r_vol_yr	Mean annual volume	m^3s^{-1}
r_height_yr	Mean annual runoff depth	mm

3 Data and code availability

The GRDC-Caravan extension dataset is publicly available on Zenodo: https://zenodo.org/records/14006282 (Färber et al., 2024). The original code to produce Caravan extensions is available here: https://github.com/kratzert/Caravan (Kratzert 2024). The R package 'GRDCFlowTools' can be provided on request.

4 Outlook

This paper introduces a Caravan extension extending the original data set in space, but also in time. Although some countries are already covered in the base Caravan dataset, GRDC stations for the same countries such as the USA, Canada, Brazil, and others have been added because the GRDC time series have brought an additional 70 years of discharge data for over 700 stations. We show that by using the Caravan platform, flow data from any set of catchments can be augmented using hydrometeorological time series and catchment attributes and formatted in a way that enables its immediate use as part of the Caravan dataset. This benefits the community and because the extraction process is automated, additional catchments can easily be added.

Currently, most of the data in the Global Runoff Database is published under a data policy which requires identified access, prohibiting redistribution and commercial use. Without the 25 WMO member states who provided their data under an open access license, the publication of the GRDC caravan extension would not have been possible. As GRDC is continuously receiving new data from NMHS, it is intended that this dataset will be updated. We hope that in the future more and more member states will make their data available under open licenses so that it can be included in future versions of the extension. Future versions of the extension will also benefit from upgrades of the Caravan dataset, for instance the planned inclusion of weather forecasts/hindcasts. Overall, we see these future developments as two-way exchanges between this extension and the wider Caravan dataset: this extension extends Caravan and, at the same time, will benefit from future Caravan improvements. Extending the dataset in space and time are just two ways to contribute to Caravan, and here we invite members of the community to imagine and share other extensions, for instance including time series for new variables or new landscape attributes.

Author contributions

HP, FK, NA, GS, SM and UL were responsible for the design and processing of the dataset; CF, HP, FK and NA organized and wrote the manuscript. All authors discussed the results and contributed to the final paper.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

GRDC thanks the World Meteorological Organization (WMO) for providing support and the Federal Institute of Hydrology, Germany (BfG) for funding. The NMHS are acknowledged for providing their valuable data sets.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293-5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

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

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrol. Earth Syst. Sci., 22, 5817-5846, 10.5194/hess-22-5817-2018, 2018.

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., and Poulin, A.: 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.

Bloomfield, J. P., Gong, M., Marchant, B. P., Coxon, G., and Addor, N.: How is Baseflow Index (BFI) impacted by water resource management practices?, Hydrol. Earth Syst. Sci., 25, 5355-5379, https://doi.org/10.5194/hess-25-5355-2021, 2021.

Casado Rodríguez, J.: CAMELS-ES: Catchment Attributes and Meteorology for Large-Sample Studies – Spain (1.0.2) [dataset], https://doi.org/10.5281/zenodo.8428374, 2023.

Chagas, V. B. P., Chaffe, P. L. B., and Blöschl, G.: Climate and land management accelerate the Brazilian water cycle, Nature Communications, 13, 5136, https://doi.org/10.1038/s41467-022-32580-x, 2022.

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 Syst. Sci. Data, 12, 2075-2096, 10.5194/essd-12-2075-2020, 2020.

Clerc-Schwarzenbach, F. M., Selleri, G., Neri, M., Toth, E., van Meerveld, I., and Seibert, J.: HESS Opinions: A few camels or a whole caravan?, EGUsphere, 2024, 1-29, 10.5194/egusphere-2024-864, 2024.

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 Syst. Sci. Data, 12, 2459-2483, 10.5194/essd-12-2459-2020, 2020.

Csardi, G. and Nepusz, T.: The igraph software package for complex network research, InterJournal, Complex Systems, 1695, https://igraph.org, 2006.

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata, Earth Syst. Sci. Data, 10, 765-785, https://doi.org/10.5194/essd-10-765-2018, 2018.

Dorigo, W., Dietrich, S., Aires, F., Brocca, L., Carter, S., Cretaux, J.-F., Dunkerley, D., Enomoto, H., Forsberg, R., Güntner, A., Hegglin, M. I., Hollmann, R., Hurst, D. F., Johannessen, J. A., Kummerow, C., Lee, T., Luojus, K., Looser, U., Miralles, D. G., Pellet, V., Recknagel, T., Vargas, C. R., Schneider, U., Schoeneich, P., Schröder, M., Tapper, N., Vuglinsky, V., Wagner, W., Yu, L., Zappa, L., Zemp, M., and Aich, V.: Closing the Water Cycle from Observations across Scales: Where Do We Stand?, Bulletin of the American Meteorological Society, 102, E1897-E1935, https://doi.org/10.1175/BAMS-D-19-0316.1, 2021.

Färber, C., Plessow, H., Kratzert, F., Addor, N., Shalev, G., and Looser, U.: GRDC-Caravan: extending the original dataset with data from the Global Runoff Data Centre (0.3) [dataset], https://doi.org/10.5281/zenodo.14006282, 2024.

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 Syst. Sci. Data, 13, 3847-3867, 10.5194/essd-13-3847-2021, 2021.

GCOS: The Status of the Global Climate Observing System 2021: The GCOS Status Report (GCOS-240), WMO, Geneva, 384 pp., 2021.

GEO: The GEOSS Water Strategy: From Observations to Decisions, Japan Aerospace Exploration Agency, Tokyo, 276 pp., 2014.

GRDC: Policy guidelines for the dissemination of data: https://grdc.bafg.de/about/data_policy/, last access: 19 September 2024, 2024a.

GRDC: Data Portal: https://grdc.bafg.de/data/data_portal/, last access: 19 September 2024, 2024b.

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment, Earth Syst. Sci. Data, 10, 787-804, https://doi.org/10.5194/essd-10-787-2018, 2018.

Heberger, M.: delineator.py: Fast, accurate watershed delineation using hybrid vector- and raster-based methods and data from MERIT-Hydro (Version 1.0) [dataset], https://doi.org/10.5281/zenodo.7314287, 2021.

Helgason, H. B. and Nijssen, B.: LamaH-Ice: LArge-SaMple DAta for Hydrology and Environmental Sciences for Iceland [dataset], https://doi.org/10.4211/hs.86117a5f36cc4b7c90a5d54e18161c91, 2024.

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.: Catchment attributes and hydro-meteorological time series for large-sample studies across hydrologic Switzerland (CAMELS-CH) (0.8) [dataset], https://doi.org/10.5281/zenodo.10354485, 2023.

Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe, Earth Syst. Sci. Data, 13, 4529-4565, 10.5194/essd-13-4529-2021, 2021.

Koch, J.: Caravan extension Denmark - Danish dataset for large-sample hydrology (v_05) [dataset], https://doi.org/10.5281/zenodo.7962379, 2022.

Kratzert, F.: Caravan - A global community dataset for large-sample hydrology: https://github.com/kratzert/Caravan, last access: 19 September 2024.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resources Research, 55, 11344-11354, https://doi.org/10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089-5110, https://doi.org/10.5194/hess-23-5089-2019, 2019b.

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.

Lawford, R., Unninayar, S., Huffman, G. J., Grabs, W., Gutiérrez, A., Ishida-Watanabe, C., and Koike, T.: Implementing the GEOSS water strategy: from observations to decisions, International Journal of Digital Earth, 16, 1439-1468, https://doi.org/10.1080/17538947.2023.2202420, 2023.

Lehner, B.: Derivation of watershed boundaries for GRDC gauging stations based on the hydrosheds drainage network, GRDC Report Series, 41, Koblenz, Germany, 18 pp., 2012.

Lehner, B., Linke, S., and Thieme, M.: HydroATLAS version 1.0 figshare [dataset], https://doi.org/10.6084/m9.figshare.9890531.v1, 2019.

Lehner, B., Verdin, K., and Jarvis, A.: New Global Hydrography Derived From Spaceborne Elevation Data, EOS, 89, 93-94, https://doi.org/10.1029/2008EO100001, 2008.

Lehner, B., Messager, M. L., Korver, M. C., and Linke, S.: Global hydro-environmental lake characteristics at high spatial resolution, Scientific Data, 9, 351, https://doi.org/10.1038/s41597-022-01425-z, 2022.

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.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, Scientific Data, 6, 283, https://doi.org/10.1038/s41597-019-0300-6, 2019.

Metzger, M. J., Bunce, R. G. H., Jongman, R. H. G., Sayre, R., Trabucco, A., and Zomer, R.: A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring, Global Ecology and Biogeography, 22, 630-638, https://doi.org/10.1111/geb.12022, 2013.

Morin, E.: Caravan extension Israel - Israel dataset for large-sample hydrology [dataset], https://doi.org/10.5281/zenodo.12760798, 2024.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J. N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349-4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

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, Hydrol. Earth Syst. Sci., 19, 209-223, https://doi.org/10.5194/hess-19-209-2015, 2015.

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

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.