Data description to 'GRDC-Caravan: extending the original dataset with data from the Global Runoff Data Centre'

2024-10-29

This document briefly describes the content and document structure of the GRDC-Caravan extension dataset. The official publication is currently in submission to ESSD.

Summary

Large-sample datasets are essential in hydrological science to support modelling studies and global assessments. This dataset is an extension to Caravan, a global community dataset of meteorological forcing data, catchment attributes, and discharge data for catchments around the world (Kratzert et al. 2023[1]).

The extension includes a subset of those hydrological discharge data and station-based watersheds from the Global Runoff Data Centre (GRDC), which are covered by an open data policy (Attribution 4.0 International; CC BY 4.0). In total, the dataset covers stations from 5356 catchments and 25 countries worldwide with a time series record from 1950 – 2023.

GRDC is an international data centre operating under the auspices of the World Meteorological Organization (WMO) at the German Federal Institute of Hydrology (BfG). Established in 1988, it holds the most substantive collection of quality assured river discharge data worldwide. Primary providers of river discharge data and associated metadata are the National Hydrological and Hydro-Meteorological Services of WMO Member States.

[1]Kratzert, F., Nearing, G., Addor, N. et al. Caravan - A global community dataset for large-sample hydrology. Sci Data 10, 61 (2023). https://doi.org/10.1038/s41597-023-01975-w

Dataset structure

The dataset is provided in the following two file formats:

1. `caravan-grdc-extension-csv.tar.gz`: provides the time series data as comma-separated text files (CSV) (downloadable as 9.6 GB zip archive)
2. `caravan-grdc-extension-nc.tar.gz`: provides the time series data in the Network Common Data Form (NetCDF) (downloadable as 8 GB zip archive)

The data in both versions are identical, users can choose if they require the time series data in CSV or NetCDF format.

Both datasets are organized in the following subfolder structure, following the dataset structure of the base Caravan dataset:

- The *attributes* folder contains a subfolder with four csv (comma-separated values) files. The file 'attributes_hydroatlas_grdc.csv' contains attributes derived from HydroATLAS and the file 'attributes_caravan_grdc.csv' contains climate indices derived from ERA5-Land. The other two files are specific for the GRDC extension and include one file with attributes from GRDC ('attributes_other_grdc.csv') and one with the original national station IDs ('national_station_ids.csv'). The first column in all attributes file is called 'gauge_id' and contains a unique station identifier of the source dataset (grdc) and the station id as defined in the original source dataset.
- The *shapefiles* folder contains a subfolder with a shapefile with the catchment boundaries of each station within the dataset. This shapefile was used to derive the catchment attributes and ERA5-Land time series data. Each polygon in a given shapefile has a field 'gauge_id' that contains the unique station identifier.
- The *timeseries* folder contains one subfolder (*csv* or *netcdf*, depending on the chosen file format). Within the subdirectory, there is one file (either *csv* or *netcdf*) per basin, containing all time series data (meteorological forcings, state variables, and streamflow). The netCDF files also contain metadata information, including physical units, timezones, and information on the data sources.
- The *licenses* folder contains license information of all data included in the extension.

Technical details

Units of time series data

Variable name	Description	Aggregation	Unit
snow_depth_water_equivalent	Snow-Water Equivalent	Daily min, max, mean	mm
surface_net_solar_radiation	Surface net solar radiation	Daily min, max, mean	W/m2
surface_net_thermal_radiation	Surface net thermal radiation	Daily min, max, mean	W/m2
surface_pressure	Surface pressure	Daily min, max, mean	kPa
temperature_2m	2m air temperature	Daily min, max, mean	°C
u_component_of_wind_10m	U-component of wind at 10m	Daily min, max, mean	m/s
v_component_of_wind_10m	V-component of wind at 10m	Daily min, max, mean	m/s
volumetric_soil_water_layer_1	volumetric soil water layer 1 (0-7cm)	Daily min, max, mean	m3/m3
volumetric_soil_water_layer_2	volumetric soil water layer 2 (7-28cm)	Daily min, max, mean	m3/m3
volumetric_soil_water_layer_3	volumetric soil water layer 3 (28-100cm)	Daily min, max, mean	m3/m3
volumetric_soil_water_layer_4	volumetric soil water layer 4 (100-289cm)	Daily min, max, mean	m3/m3
total_precipitation	Total precipitation	Daily sum	mm
potential_evaporation	Potential Evapotranspiration	Daily sum	mm
streamflow	Observed streamflow	Daily min, max, mean	mm/d

Catchment attributes

Refer to the Caravan paper for a detailed list of all static catchment attributes. Generally, there are two different sets of catchment attributes that are shared in two different files:

1. `attributes_hydroatlas_grdc.csv`: Attributes derived from the HydroATLAS dataset. Refer to the "BasinATLAS Catalogue" of HydroATLAS (see here) for an explanation of the features.
2. `attributes_caravan_grdc.csv`: Attributes (climate indices) derived from ERA5-Land timeseries for the Caravan dataset. See Table 4 in the Caravan paper for details.
3. `attributes_other_grdc.csv`: Metadata information, such as station name, gauge latitude and logitude, country and area. See Table 5 in the Caravan paper for details.

Additional attributes from GRDC

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 since	year
d_yrs	Length of time series of daily data	-
d_miss	Percentage of missing values	-
lta_discharge	Long-term average discharge	m^3s^{-1}
r_vol_yr	Mean annual volume	m^3
r_height_yr	Mean annual runoff depth	mm

Time zones

All data in Caravan are in the local time of the corresponding catchment. We ignore any possible Daylight Saving Time, i.e., data for a given location is always in non-DST time, regardless of the date.

License

See `licenses/README.md` for details on the license of Caravan as well as all source datasets.

How to cite

The GRDC-Caravan extension paper is currently in submission. If you already use Caravan in your research papers, please cite the dataset using the following reference:

Färber, C., Plessow, H., Kratzert, F., Addor, N., Shalev, G., & Looser, U. (2023). GRDC-Caravan: extending the original dataset with data from the Global Runoff Data Centre (0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8425587