



Spatially resolved meteorological and ancillary data in Central Europe for rainfall streamflow modeling

Marc Aurel Vischer, Noelia Otero, and Jackie Ma

Fraunhofer Heinrich-Hertz Institute, Applied Machine Learning Group, 10587 Berlin, Germany

Correspondence: Marc Aurel Vischer (marc.aurel.vischer@hhi.fraunhofer.de) and Jackie Ma (jackie.ma@hhi.fraunhofer.de)

Received: 12 September 2025 – Discussion started: 11 November 2025

Revised: 5 February 2026 – Accepted: 30 March 2026 – Published: 8 May 2026

Abstract. We present a gridded dataset for rainfall streamflow modeling that is fully spatially resolved and covers five complete river basins in central Europe: upper Danube, Elbe, Oder, Rhine, and Weser (<https://doi.org/10.4211/hs.d7f2cbb587ab4a75ac7987854e8f62ca>, Vischer et al., 2025a). We compiled meteorological forcings and a variety of ancillary information on soil, rock, land cover, and orography. The data is harmonized to a regular 9 km × 9 km grid, temporal resolution is daily from 1980 to 2024. We also provide code to further combine our dataset with publicly available river discharge data for end-to-end rainfall streamflow modeling. We have used this data to demonstrate how neural network-driven hydrological modeling can be taken beyond lumped catchments, and want to facilitate direct comparisons between different model types.

1 Introduction

In recent years, a substantial number of rainfall streamflow datasets were released that follow the example of the popular CAMELS dataset (Newman et al., 2015; Addor et al., 2017). They cover Chile (Alvarez-Garreton et al., 2018), Great Britain (Coxon et al., 2020), Brazil (Chagas et al., 2020), Australia (Fowler et al., 2021), the upper Danube basin (Klingler et al., 2021), France (Delaigue et al., 2022), Switzerland (Höge et al., 2023), Denmark (Liu et al., 2024) and Germany (Loritz et al., 2024). There is also a world-wide version of CAMELS, CARAVAN (Kratzert et al., 2023). These datasets bundle a range of data sources and harmonize them to a readily ingestible, common spatio-temporal data format. Besides meteorological variables, they contain additional static information such as land cover, soil and bedrock type and orographic features. While suitable for a range of hydrological modeling approaches, these publications specifically facilitated a surge in popularity of neural network models for rainfall streamflow modeling. For example, Kratzert et al. (2019); Nearing et al. (2024) have shown that neural network models are particularly suited to learn from such multi-variate, large-scale data.

A common downside of the above-mentioned datasets however is that they aggregate (“lump”) each variable within a catchment to a single value. By doing so, all information about spatial variability is lost: A pattern of soil types might be reduced to the most prevalent one, or a range of different amounts of precipitation over a large area might be averaged to a single, unexpressive average value. This reduction of information is unnecessary and counter-intuitive, especially for large catchments or catchments with high spatial variability. The principle advantage of spatially resolved inputs is that they enable the model to capture spatial covariance among different variables, e.g. the interacting effects of soil sealing or steepness of terrain and a torrential rainfall. Physical models, still the standard model type in active operation, resolve their equations on such a grid for exactly this reason, but neural network training also benefit from vast amounts of data. Additionally, as each grid cell contains a complete, self-contained set of meteorological and ancillary variables, they can be processed independently. Recent advances in memory parallelism can leverage this property and make large-scale processing of such data without prior aggregation practically feasible: in Vischer et al. (2025b), we show that a neural network model is capable of efficiently handling this

large amount of data. However, parallelization is just as applicable for physical or conceptual modeling approaches. Recently, Kraft et al. (2025) demonstrated this by using a combined neural network, conceptual and physical approach on spatially resolved data from Switzerland to significantly outperform a lumped baseline. Rakhymbek et al. (2025) report similar results with a neural network model trained on two snow-driven basins in Kazakhstan and the USA. Both studies highlight the importance of representing spatial detail. Gauch et al. (2024) and Gauch et al. (2025) discussed general advantages and applicability of semi-distributed models for future world-wide flood-prediction models at recent conferences, further testifying to the relevance of providing datasets suitable for this research direction.

2 Methods

The study area of the dataset covers 5 entire basins in central Europe, namely the upper reaches of the Danube (until Bratislava), Elbe, Oder, Rhine and Weser. It is contiguous, 570.592 km² large and spans 10 countries. The temporal coverage ranges from 1 January 1980 to 31 December 2024. We bundle 6 spatiotemporal (“dynamic”) meteorological features with 46 static (“ancillary”) features: 3 hydro-geological features, 16 land cover features, 19 soil features and 8 orographical features. We based our choice of which kind of dynamic and ancillary information to include on the work of Addor et al. (2017) and Kratzert et al. (2019) to allow for maximum comparability with recent hydrological literature in general and neural network-based literature in particular. The dataset consists of data derived from a variety of publicly available sources – no new data was recorded. Our contribution consists in collecting the data and harmonizing it to a common grid: We provide a comprehensive, peer-reviewed, publicly licensed dataset that integrates well with already existing datasets while supporting novel research questions regarding spatial heterogeneity. Figure 1 provides an overview of the study area, common grid and types of variables. The remainder of this section first describes each data modality. Original data sources are listed in Table A1, detailed lists of all dynamic and ancillary features that we derived to compile this dataset can be found in Tables A2 and A3. We then describe how the different spatial data sources were harmonized to a common grid. Along with the data, we release all scripts for processing the raw source data into the dataset. This allows users to both verify and adapt our data aggregation pipeline. We also provide an additional script that combines the dataset presented here with river discharge data, after manual download from the original provider, the Global Runoff Data Center (GRDC) (https://grdc.bafg.de/data/data_portal/, last access: 11 September 2025). In the preprocessing code linked below, we also show that our study area is covered densely and uniformly with river gauging stations. As there are much fewer

stations in the lower Danube basin, we decided to only include the upper part in order to eliminate this source of sampling bias, which can be a major concern not only for data-driven approaches. The discharge time series come at daily resolution, which is the reason that we provide our temporal features in daily resolution as well. This data can serve as targets for end-to-end training in data-driven rainfall streamflow modeling, such as in our study (Vischer et al., 2025b).

2.1 Meteorological Forcings

The meteorological forcings in our study were derived from the ERA5-Land dataset¹ (Muñoz Sabater, 2019; Copernicus Climate Change Service, 2022). Temporal aggregation from hourly to daily resolution was achieved differently for each variable: The temperature two meters above surface was aggregated by calculating minimum, mean and maximum values. Potential evapotranspiration was summed. Precipitation is provided in ERA5-Land as sub-daily values, meaning that the daily total sum corresponds to the value stored at 24:00. We added a measure of variability of precipitation by taking the variance over the increment at every hourly time step. Due to the good maintenance of the ERA5 dataset, there were no missing values in the temporal data, hence no interpolation was necessary on our part. Table A2 provides a detailed list of all dynamic variables.

2.2 Ancillary Data

Hydrogeological properties were derived from the International Hydrogeological Map of Europe (IHME)². The original dataset features six hydrogeological classes as well as two classes for snow-ice-fields and inland water bodies. The six classes represent the productivity of rock type, which indicates how easily water can dissipate through the bedrock. Classes are ordinal in that they are sorted by the corresponding productivity in ascending order. This allows us to take a non-rigorously defined but nonetheless informative average over the classes’ proportions within each grid cell. We concatenate this productivity score with the binary categorical classes for snow-ice-fields and inland water bodies, each rep-

¹The dataset was downloaded from the Copernicus Climate Change Service (2022) (<https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview>, last access: 11 September 2025). The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

²IHME1500 – Internationale Hydrogeologische Karte von Europa 1 : 1 500 000, version 1.2 (https://www.bgr.bund.de/EN/Themen/Wasser/Projekte/laufend/Beratung/Ihme1500/ihme1500_projektbeschr_en.html, last access: 11 September 2025) © Bundesanstalt für Geowissenschaft und Rohstoffe, 2022 (Günther and Duscher, 2019)

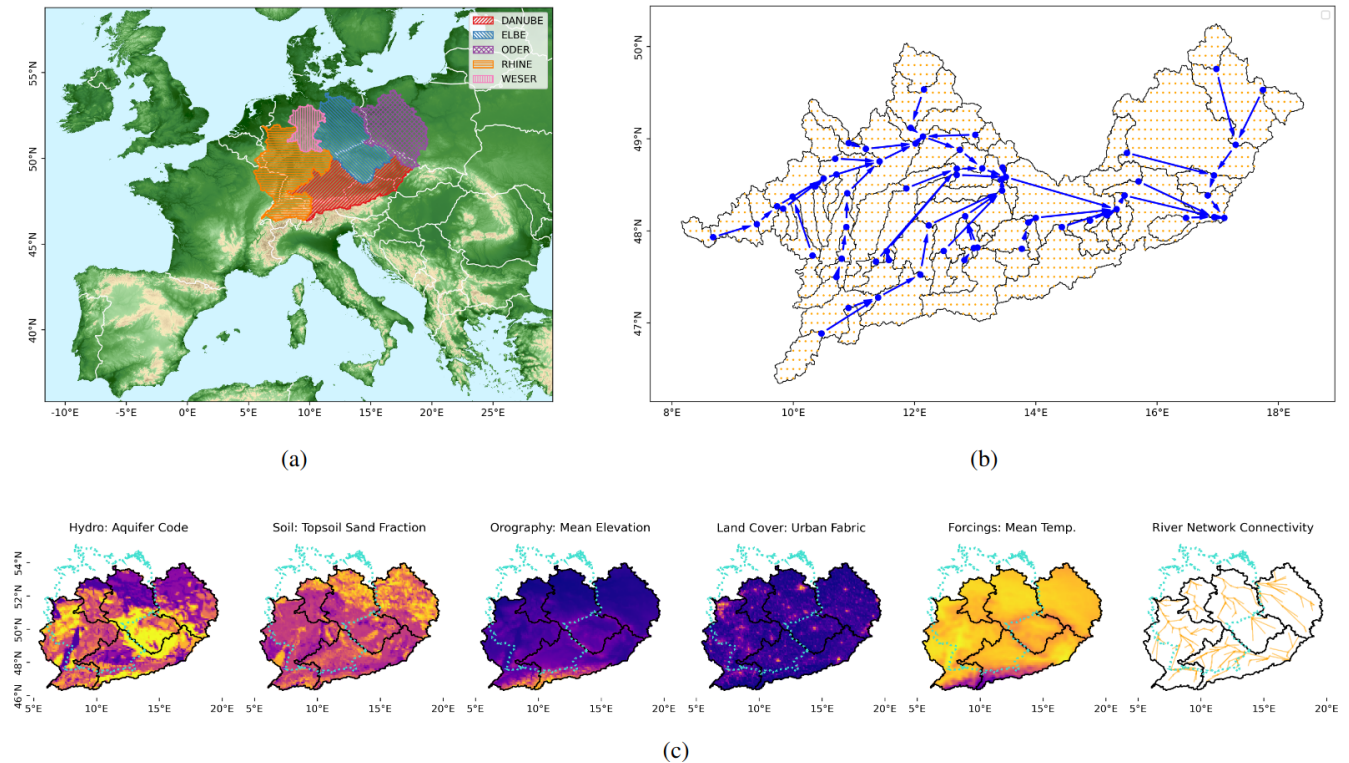


Figure 1. Overview of study area, input grid and data types. **(a)** Study Area: The study area comprises 5 basins that cover a contiguous area in central Europe. **(b)** Input Grid and Station Network in Upper Danube Basin: Cells of input grid (orange) for Upper Danube basin. Catchment boundaries (black) are overlaid with corresponding stations (blue), as well as connecting arrows representing the station connectivity network. Code to reproduce the river network is released together with this paper. **(c)** Input Types: Visualizations for one example feature of each type of input. Basin outlines (black) and borders of Germany (turquoise) are plotted for reference. Orography in panel **(a)** was adapted from the European Space Agency’s Copernicus Global 90 m DEM (GLO-90, doi: <https://doi.org/10.5270/ESA-c5d3d65>, ESA, 2022) © EuroGeographics for the administrative boundaries in panel **(a)** and **(c)**. Watershed boundaries in panels **(a)**, **(b)**, and **(c)** were taken from the Global Runoff Data Center (<https://portal.grdc.bafg.de/applications/public.html>, last access: 11 September 2025).

resented by a ratio of prevalence of this type of binary class within the grid cell.

Land Cover information was obtained from the Corine Land Cover Map³ (CLC). This dataset classifies land cover at three different levels of detail, with increasingly differentiated (sub)classes. We decided to use the second level, which containing 16 classes in total. Similarly to the procedure applied to the hydrogeological properties, we calculated a distributional vector representing the proportion of a given class covering the grid cell.

³Corine Land Cover Map, version 2012 (<https://land.copernicus.eu/en/products/corine-land-cover/clc-2012>, last access: 11 September 2025). Generated using European Union’s Copernicus Land Monitoring Service information; <https://doi.org/10.2909/916c0ee7-9711-4996-9876-95ea45ce1d27> (Copernicus Land Monitoring Service, 2020). The Corine Land Cover Map data was created with funding by the European union. It was adapted and modified by the authors.

Soil type information was obtained from the dataset European Soil Database Derived Data⁴ (Hiederer, 2013a, b). This dataset features 17 different physical properties, separately for top soil and lower soil. We calculate the average value of each feature within a grid cell.

Orographic information was derived from the European Union Digital Elevation Map⁵ (EU-DEM). Elevation was av-

⁴European Soil Database Derived Data (<https://esdac.jrc.ec.europa.eu/content/european-soil-database-derived-data>, last access: 11 September 2025), created by the European Soil Data Centre with funding by the European union. It was adapted and modified by the authors. The authors’ activities are not officially endorsed by the Union.

⁵European Union Digital Elevation Map, version 1.1 (<https://sdi.eea.europa.eu/catalogue/srv/api/records/d08852bc-7b5f-4835-a776-08362e2fbf4b>, last access: 11 September 2025). Generated using European Union’s Copernicus Land Monitoring Service information. The European Union Digital Elevation Map created with funding by the European union. It was adapted and modified by the authors. The authors’ activities are not officially endorsed by the Union.

eraged within each grid cell, as well as the gradient in latitudinal and longitudinal direction, and the steepness or magnitude of the two-dimensional gradient. This yielded a total of four orographic features.

Table A3 provides an detailed list of all ancillary variables in the same ordering as we just introduced, which is also the ordering in the data file.

2.3 Grid Harmonization

As a common grid format, we decided to use the grid of the ERA5-Land reanalysis dataset (Muñoz Sabater, 2019; Copernicus Climate Change Service, 2022), which covers the earth's surface with a $0.1^\circ \times 0.1^\circ$ resolution, corresponding to roughly $9 \text{ km} \times 9 \text{ km}$ for the case of our study area. ERA5-Land contains a vast number of meteorological variables and has an hourly resolution, spanning from 1950 onward. It has been widely used and is actively maintained and updated. This means the dataset we provide with this paper remains easily extendable, should a user like to e.g. include an additional meteorological variable in their experiments, extend the study area or increase the temporal resolution. Our study area comprises a total of 7169 grid cells on this grid.

The spatial data originally comes in different formats (vector or grid), projections and resolutions. All data sources were thus harmonized to the grid of ERA5-Land. The first step consisted in converting the maps to a common coordinate system. For the sake of compatibility with ERA5, we decided to use the geographic coordinate system WGS 84. Then for each map separately, polygons covering 0.1° in latitude and longitude with the grid cell at the center were extracted. For categorical maps like hydrogeology and land cover, consisting of classes such as “Artificial surfaces: Urban fabric”, the fractions covered by each class within each polygon were calculated. This results in a distributional description of class occurrence maximally conserves the original information, as no averaging or other kind of aggregation is necessary. Quantitative information, like e.g. clay content in topsoil, was however aggregated in a final step within the polygon. As a downside to this approach, note that both cases require calculating coverage areas in a geographic coordinate system. This treats the surfaces as flat instead of accounting for the earth's curvature, making the calculations imprecise. We consider this effect negligible here since the surfaces contained in the grid cells are so small that they can be safely considered approximately flat. A more severe limitation is the fact that at high latitudes, polygons defined by a given latitudinal and longitudinal extent become substantially more narrow on the side facing the pole, which also impairs the calculation of area. We could not change the polygons to counteract this effect since they are implicitly dictated by the ERA5 grid. At the moderate latitudes of our study area and especially the small polygons used in the grid, this distortion can still be considered acceptable for the sake of harmonization with ERA5, depending on the application. However, applying this

approach in polar regions for example would necessitate an intermediate step of choosing a suitable projected coordinate system for calculating the areas in order to make the distortion explicit and thus better understand its effects. Boundary effects are not an issue with this vector approach as no interpolation is required, however. All maps covered regions well exceeding the study area.

Specifically for the four map types, the hydrogeological map already came in the WGS 84 reference system. The land cover map came in vector format and LAEA coordinate system, so the polygon's coordinates could simply be calculated point-wise using the geopandas Python package (den Bossche et al., 2025) with pyproj (Snow et al., 2025) as a backend. This was followed by calculating the fraction of each class within the polygon described above. The soil map came in LAEA reference system but as a raster. We decided against interpolating as it would not contribute any additional information. Instead, we simply considered each cell in the raster as a separate polygon and calculated the fractions using the vector method described above. The situation was the same for the elevation map, with the added initial step of down-sampling the map from 25 m resolution to 500 m resolution using weighted average resampling implemented in rasterio (Mapbox, 2024).

3 Code and data availability

The dataset is available at <https://doi.org/10.4211/hs.d7f2cbb587ab4a75ac7987854e8f62ca>, (Vischer et al., 2025a). Dynamic meteorological forcing data and static ancillary data are stored in two separate NetCDF4 (Rew et al., 2006) files, “ancillary_pub.nc” and “dynamic_pub.nc”. This format allows for labeled coordinates such as latitude, longitude and date for convenient selection on spatial and temporal domains, respectively. All variables are named in a self-explanatory manner and we provide labeled metadata. See Tables A2 and A3 for a detailed reference of the included variables.

The data was processed in several Python Jupyter Notebooks (Granger and Pérez, 2021) that can be found in this repository: https://gitlab.hhi.fraunhofer.de/vischer/spatial_streamflow_dataprep, last access: 11 September 2025. The code requires Python 3.11 (Van Rossum and Drake, 2009) and is licensed under the Clear BSD licence. Additional dependencies are specified in an Anaconda environment (Anaconda, 2020) specification contained in the repository. The scripts are stand-alone and do not require further input parameters. Along with the code to process the data, we provide a script that loads the data, selects subsets and visualizes them. This can serve as a starting point for the user to interact with the data. Furthermore, we provide code to wrap all the data in a PyTorch (Ansel et al., 2024) Dataset class for further processing in a machine learning context. Since dense arrays are required for this, we pro-

vide an alternative format version of our features in the files “ancillary_paper.nc” and “dynamic_paper.nc”, where dimensions were transformed from latitude and longitude to a unique grid cell index. In this version, all features were standardized in order to suit better the requirements of neural networks.

All data sources from which we obtained the original data have been widely used across various scientific fields for years, so we assume the original data to be valid. In order to technically validate our processing steps, we feature a testing script in our repository with extensive tests and visualizations of the compiled data. We also successfully employed this dataset in training a neural network model for rainfall streamflow modeling (Vischer et al., 2025b).

4 Conclusions

Combining a variety of data sources, we provide the first spatially resolved dataset for multivariate rainfall-streamflow modeling. It covers five entire river basins and it thus particularly suited for large-scale modeling of hydrological processes. With the publication of this dataset, we hope to stimulate further development of spatially resolved, high-resolution hydrological modeling beyond the scope of lumped catchments. Suitable for neural network models as well as conceptual and physical modeling approaches, we hope this dataset will facilitate model comparison and stimulate future development in the spatially resolved domain. Using the same spatial grid as ERA5 as well as daily resolution limits its expressiveness of small scale, e.g. convective events, where higher spatial and temporal resolution would be preferable. If users decide to spatially aggregate our data and want to use derived variables like hydrological or climate signatures as input for their models, they would have to manually compute them from the raw data contained in our dataset. This is especially the case for snow-related variables, as particularly the Southeast of our study area is dominated by snowmelt dynamics. Lastly, the dataset is of only limited use for training training models that are to be deployed world-wide. We focus on a contiguous area in central Europe, which means in turn that the dataset contains only a particular subset of all hydrological dynamics that can be observed.

Appendix A: Data Origins and Detailed List of Features

Table A1. Overview of source datasets and their authors for dynamic data/meteorological forcings contained in file *forcings_pub.nc* and static/ancillary data contained in *ancillary_pub.nc*. See Tables A2 and A3 for more details on derived features. Last access for all weblinks is 11 September 2025.

| Type | Dataset | Author | Citation |
|--------------------------------|---|--|--|
| Forcings/Dynamic Inputs | | | |
| Meteorological Variables | ERA5-Land (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview) | Copernicus Climate Change Service (CCCS) (https://climate.copernicus.eu/) | Muñoz Sabater (2019) Copernicus Climate Change Service (2022) |
| Ancillary Data/Static Inputs | | | |
| Hydrogeological Properties | IHME hydrogeological map v1.2 in vector data format (https://www.bgr.bund.de/EN/Themen/Wasser/Projekte/laufend/Beratung/Ihme1500/ihme1500_projektbeschr_en.html) | German Federal Institute for Geosciences and Natural Resources (BGR) (https://www.bgr.bund.de/EN/Home/homepage_node_en.html) | Günther and Duscher (2019) |
| Land Cover | Corine Land Cover Map, version 2012 (https://land.copernicus.eu/en/products/corine-land-cover/clc-2012) | Copernicus Land Monitoring Service (CLMS) (https://land.copernicus.eu/en) | |
| Soil Type (Top and Lower Soil) | European Soil Database Derived Data (https://esdac.jrc.ec.europa.eu/content/european-soil-database-derived-data) | European Soil Data Centre (ESDAC) (https://esdac.jrc.ec.europa.eu/) | Hiederer (2013a, b) |
| Orographic | European Union Digital Elevation Map (EU-DEM), version 1.1 (https://sdi.eea.europa.eu/catalogue/srv/api/records/3473589f-0854-4601-919e-2e7dd172ff50) | Copernicus Land Monitoring Service (CLMS) (https://land.copernicus.eu/en) | |

Table A2. Overview of dynamic input features in the file *forcings_pub.nc*. Empty cells indicate that the value is identical to the one above. Each of these features is a three-dimensional array with dimensions latitude, longitude and date. Labeled coordinate indices for all dimensions are contained in the file.

| Index | Name | Feature | Origin | Aggregation | Unit |
|-------|----------|------------------------------|-----------|----------------|------|
| 00 | t2m_min | Temperature 2 m above ground | ERA5-Land | Daily Minimum | K |
| 01 | t2m_mean | | | Daily Mean | |
| 02 | t2m_max | | | Daily Maximum | |
| 03 | pev | Potential evapotranspiration | | Daily Sum | mm |
| 04 | tp_sum | Precipitation | | Daily Sum | |
| 05 | tp_var | | | Daily Variance | |

Table A3. Overview of static input features in the file *ancillary_pub.nc*. Empty cells indicate that the value is identical to the one above. Explanations of the features derived from Corine Land Cover map (CLC) and elevation map were omitted because the names are self-explanatory. Each of these features is a two-dimensional array with dimensions latitude and longitude. Labeled coordinate indices for all dimensions are contained in the file.

| Index | Name | Feature | Origin | Aggregation | Unit |
|-------|---|---|--------|--------------------|-------------------|
| 00 | IHME_AQUIF_CODE | Rock Productivity | IHME | Averaged Classes | unitless |
| 01 | IHME_INLAND_WATER | Inland Water Body | | Fraction | frac. area |
| 02 | IHME_SNOW_ICE_FIELD | Permanent Snow-Ice Field | | | |
| 03 | CLC_11_Artificial_surfaces_Urban_fabric | | CLC | | |
| 04 | CLC_12_Artificial_surfaces_Industrial_commercial_and_transport_units | | | | |
| 05 | CLC_13_Artificial_surfaces_Mine_dump_and_construction_sites | | | | |
| 06 | CLC_14_Artificial_surfaces_Artificial_non_agricultural_vegetated_areas | | | | |
| 07 | CLC_21_Agricultural_areas_Arable_land | | | | |
| 08 | CLC_22_Agricultural_areas_Permanent_crops | | | | |
| 09 | CLC_23_Agricultural_areas_Pastures | | | | |
| 10 | CLC_24_Agricultural_areas_Heterogeneous_agricultural_areas | | | | |
| 11 | CLC_31_Forest_and_seminatural_areas_Forest | | | | |
| 12 | CLC_32_Forest_and_seminatural_areas_Shrub_and_or_herbaceous_vegetation_associations | | | | |
| 13 | CLC_33_Forest_and_seminatural_areas_Open_spaces_with_little_or_no_vegetation | | | | |
| 14 | CLC_41_Wetlands_Inland_wetlands | | | | |
| 15 | CLC_42_Wetlands_Coastal_wetlands | | | | |
| 16 | CLC_51_Water_bodies_Inland_waters | | | | |
| 17 | CLC_51_Water_bodies_Marine_waters | | | | |
| 18 | CLC_No_data | | | | |
| 19 | SOIL_STU_EU_S_SILT | Subsoil: Silt Content | ESDAC | Arithmetic Mean | % |
| 20 | SOIL_STU_EU_T_SAND | Topsoil: Sand Content | | | |
| 21 | SOIL_SMU_EU_S_TAWC | Subsoil: Total Available Water Content (Pedotr. Rule) | | | mm |
| 22 | SOIL_SMU_EU_T_TAWC | Topsoil: Total Available Water Content (Pedotr. Rule) | | | |
| 23 | SOIL_STU_EU_T_BD | Topsoil: Bulk Density | | | g/cm ³ |
| 24 | SOIL_STU_EU_T_TAWC | Topsoil: Total Available Water Content (Pedotr. Function) | | | mm |
| 25 | SOIL_STU_EU_S_GRAVEL | Subsoil: Coarse Fragments | | | % |
| 26 | SOIL_STU_EU_DEPTH_ROOTS | Depth Available to Roots | | | cm |
| 27 | SOIL_STU_EU_T_GRAVEL | Topsoil: Coarse Fragments | | | % |
| 28 | SOIL_STU_EU_S_TEXT_CLS | Subsoil: Texture Class | | | unitless |
| 29 | SOIL_STU_EU_T_OC | Topsoil: Organic Content | | | % |
| 30 | SOIL_STU_EU_S_SAND | Subsoil: Sand Content | | | |
| 31 | SOIL_STU_EU_T_CLAY | Topsoil: Clay Content | | | |
| 32 | SOIL_STU_EU_T_TEXT_CLS | Topsoil: Texture Class | | | unitless |
| 33 | SOIL_STU_EU_T_SILT | Topsoil: Silt Content | | | % |
| 34 | SOIL_STU_EU_S_BD | Subsoil: Bulk Density | | | g/cm ³ |
| 35 | SOIL_STU_EU_S_TAWC | Subsoil: Total Available Water Content (Pedotr. Function) | | | mm |
| 36 | SOIL_STU_EU_S_OC | Subsoil: Organic Carbon Content | | | % |
| 37 | SOIL_STU_EU_S_CLAY | Subsoil: Clay Content | | | % |
| 38 | DEM_elevation_mean | | EU-DEM | | m |
| 39 | DEM_grad_x_mean | | | | 1/m |
| 40 | DEM_grad_y_mean | | | | |
| 41 | DEM_steepness_mean | | | | 1/m ² |
| 42 | DEM_elevation_std | | | Standard Deviation | unitless |
| 43 | DEM_grad_x_std | | | | |
| 44 | DEM_grad_y_std | | | | |
| 45 | DEM_steepness_std | | | | |

Appendix B: Comparison with Related Datasets

Table B1. This table compares the variables contained in our dataset to those contained in similar datasets. We matched the raw variables contained in the CAMELS dataset (Addor et al., 2017 and Newman et al., 2015) as precisely as data availability in our study area permits. This means that e.g. the classes for describing land cover type may vary, although the kind of information is the same. Our dataset also matches other established datasets in this domain rather closely in terms of selection of variables, namely CARAVAN (Kratzert et al., 2023), CAMELS-DE (Loritz et al., 2024) and CAMELS-GB (Coxon et al., 2020). The only substantial difference to all these datasets is that we opted not to include derived climate and hydrological signatures of basins, since any aggregation of our data is optional and would depend on the application. Arguably, since spatially resolved data is more abundant and detailed, it might render “summary statistics” of entire basins less relevant to begin with. In any case, signatures can be readily calculated from the raw data contained in our dataset according to the user’s preference. Throughout the table, all listed temporal variables across all datasets are aggregated daily, so “mean” temperature refers to daily mean etc.

| | | Our Dataset | CAMELS | CARAVAN | CAMELS-DE | CAMELS-GB |
|----------------------------|---|-------------|--------|---------|-----------|-----------|
| Dynamic/meteo. variables | Temperature minimum | ✓ | ✓ | ✓ | ✓ | |
| | Temperature mean | ✓ | | ✓ | ✓ | ✓ |
| | Temperature maximum | ✓ | ✓ | ✓ | ✓ | |
| | Dew point temperature min, mean, max | | | ✓ | | |
| | Potential evapotranspiration | ✓ | | ✓ | ✓ | ✓ |
| | Pot. evapot. (rain corrected) | | | | | ✓ |
| | Daily precipitation sum | ✓ | ✓ | ✓ | | ✓ |
| | Daily precipitation variance | ✓ | | | | |
| | Daily precip. min, mean, median, max, std | | | | ✓ | |
| | Wind components N- and E-ward, min, mean, max | | | ✓ | | |
| | Windspeed | | | | | ✓ |
| | Humidity mean | | ✓ | | ✓ | ✓ |
| | Humidity minimum | | | | ✓ | |
| | Humidity maximum | | | | ✓ | |
| | Shortwave radiation | | ✓ | ✓ | | ✓ |
| | Shortwave radiation min, max | | | ✓ | | |
| | Longwave radiation | | | ✓ | | ✓ |
| | Longwave radiation min, max | | | ✓ | | |
| | Net surface radiation min, mean, max | | | ✓ | ✓ | |
| | Net surface radiation median, std | | | | ✓ | |
| | Surface pressure min, mean, max | | | ✓ | | |
| | Day length | | | ✓ | | |
| | Snow water equivalent | | | ✓ | | |
| Ancillary/static variables | Hydrogeology | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Land Cover | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Soil Type | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Orography | ✓ | ✓ | ✓ | ✓ | ✓ |
| Basin signatures | Hydrological signatures | | ✓ | ✓ | ✓ | ✓ |
| | Climate signatures | | ✓ | ✓ | ✓ | |
| | Gauging station properties | | ✓ | ✓ | ✓ | ✓ |
| | Human influence attributes | | | ✓ | ✓ | ✓ |
| | Simulated hydrologic time series (model output) | | | | ✓ | |
| River gauge data | Catchment discharge | | | | ✓ | ✓ |
| | Catchment-specific discharge | | | | ✓ | ✓ |
| | Water level | | | | ✓ | |

Author contributions. M.A.V. compiled the data with crucial suggestions from N.O., processed the data, and wrote the manuscript with significant contributions from J.M. All authors reviewed the manuscript.

Interactive computing environment

The Jupyter Notebooks used for data pre-processing and dataset preparation are available at https://gitlab.hhi.fraunhofer.de/vischer/spatial_streamflow_dataprep/-/tree/master/deployment (last access: 11 September 2025). The repository includes a conda environment file (`environment_preproc.yml`) that specifies all software dependencies, allowing readers to reproduce the data preparation workflow by recreating the environment using the Conda package manager and following the instructions provided in the repository README.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Financial support. This work was supported by the Federal Ministry for Economic Affairs and Climate Action (BMWK) as grant DAKI-FWS (01MK21009A), and by the European Union's Horizon Europe research and innovation program (EU Horizon Europe) project MedEWSa under grant agreement no. 101121192.

Review statement. This paper was edited by Alexander Gelfan and reviewed by two anonymous referees.

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.
- Alvarez-Garretón, 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, <https://doi.org/10.5194/hess-22-5817-2018>, 2018.
- Anaconda: Anaconda Software Distribution, Computer Software, 2020.
- Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C. K., Maher, B., Pan, Y., Puhersch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Zhang, S., Suo, M., Tillet, P., Zhao, X., Wang, E., Zhou, K., Zou, R., Wang, X., Mathews, A., Wen, W., Chanan, G., Wu, P., and Chintala, S.: PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation, in: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Vol. 2, ACM, La Jolla CA USA, 929–947, ISBN 979-8-4007-0385-0, <https://doi.org/10.1145/3620665.3640366>, 2024.
- 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, <https://doi.org/10.5194/essd-12-2075-2020>, 2020.
- Copernicus Land Monitoring Service: CORINE Land Cover 2012 (vector), Europe, version 2020_20u1, European Environment Agency (EEA), <https://doi.org/10.2909/916c0ee7-9711-4996-9876-95ea45ce1d27>, 2020.
- Copernicus Climate Change Service: ERA5-Land Hourly Data from 1950 to Present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], <https://doi.org/10.24381/Cds.E2161bac>, 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 Syst. Sci. Data*, 12, 2459–2483, <https://doi.org/10.5194/essd-12-2459-2020>, 2020.
- Delaigue, O., Brigode, P., Andréassian, V., Perrin, C., Etchevers, P., Soubeyroux, J.-M., Janet, B., and Addor, N.: CAMELS-FR: A large sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking, IAHS-AISH Scientific Assembly 2022, Montpellier, France, 29 May–3 Jun 2022, IAHS2022-521, <https://doi.org/10.5194/iahs2022-521>, 2022.
- den Bossche, J. V., Jordahl, K., Fleischmann, M., Richards, M., McBride, J., Wasserman, J., Badaracco, A. G., Snow, A. D., Roggemans, P., Ward, B., Tratner, J., Gerard, J., Perry, M., Taves, M., Hjelle, G. A., carsonfarmer, Tan, N. Y., Bell, R., ter Hoeven, E., Caria, G., Cochran, M. D., rraymondgh, Culbertson, L., Bartos, M., Chai, C. P., Eubank, N., sangarshanan, Flavin, J., and Rey, S.: Geopandas/Geopandas: V1.1.2, Zenodo [code], <https://doi.org/10.5281/zenodo.18024156>, 2025.
- European Space Agency (ESA): Copernicus DEM Global and European Digital Elevation Model (COP-DEM), Copernicus Data Space Ecosystem, <https://doi.org/10.5270/ESA-c5d3d65>, 2022.
- 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, <https://doi.org/10.5194/essd-13-3847-2021>, 2021.
- Gauch, M., Kratzert, F., Dube, V., Gilon, O., Klotz, D., Metzger, A., Nearing, G., Ofori, F., Shalev, G., Shenzi, S., Tekalgin,

- T., Weitzner, D., Zlydenko, O., and Cohen, D.: Deep Learning for Spatially Distributed Rainfall–Runoff Modeling, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-8899, <https://doi.org/10.5194/egusphere-egu24-8899>, 2024.
- Gauch, M., Kratzert, F., Metzger, A., Shenzi, S., Klotz, D., Cohen, D., and Gilon, O.: Semi-Distributed Hydrological Modeling Based on Deep Learning at Scale, EMS Annual Meeting 2025, Ljubljana, Slovenia, 7–12 Sep 2025, EMS2025-31, <https://doi.org/10.5194/ems2025-31>, 2025.
- Granger, B. E. and Pérez, F.: Jupyter: Thinking and Storytelling With Code and Data, *Comput. Sci. Eng.*, 23, 7–14, <https://doi.org/10.1109/MCSE.2021.3059263>, 2021.
- Günther, A. and Duscher, K.: Extended Vector Data of the International Hydrogeological Map of Europe 1: 1,500,000 (Version IHME1500 v1. 2), Federal Institute for Geosciences and Natural Resources (BGR), Hannover, Berlin, Germany, 2019.
- Hiederer, R.: Mapping Soil Typologies: Spatial Decision Support Applied to the European Soil Database, Publications Office of the European Union 127, <https://doi.org/10.2788/87286>, 2013a.
- Hiederer, R.: Mapping Soil Properties for Europe: Spatial Representation of Soil Database Attributes., EUR26082EN scientific and technical research series 47, <https://doi.org/10.2788/94128>, 2013b.
- 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, *Earth Syst. Sci. Data*, 15, 5755–5784, <https://doi.org/10.5194/essd-15-5755-2023>, 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, <https://doi.org/10.5194/essd-13-4529-2021>, 2021.
- Kraft, B., Kauzlaric, M., Aeberhard, W. H., Zappa, M., and Gudmundsson, L.: DROP: A Scalable Deep Learning Approach for Runoff Simulation and River Routing, <https://www.authorea.com/doi/full/10.22541/au.176410929.91946608/v1>, 2025.
- 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 Resour. Res.*, 55, 11344–11354, <https://doi.org/10.1029/2019WR026065>, 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, *Sci. Data*, 10, 61, <https://doi.org/10.1038/s41597-023-01975-w>, 2023.
- Liu, J., Koch, J., Stisen, S., Troldborg, L., Højberg, A. L., Thodsen, H., Hansen, M. F. T., and Schneider, R. J. M.: CAMELS-DK: hydrometeorological time series and landscape attributes for 3330 Danish catchments with streamflow observations from 304 gauged stations, *Earth Syst. Sci. Data*, 17, 1551–1572, <https://doi.org/10.5194/essd-17-1551-2025>, 2025.
- Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S. K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., and Tarasova, L.: CAMELS-DE: hydro-meteorological time series and attributes for 1582 catchments in Germany, *Earth Syst. Sci. Data*, 16, 5625–5642, <https://doi.org/10.5194/essd-16-5625-2024>, 2024.
- Mapbox: Rasterio v1.4.3, Mapbox, <https://github.com/rasterio/rasterio> (last access: 30 April 2026), 2024.
- Muñoz Sabater, J.: ERA5-Land Hourly Data from 1950 to Present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], <https://doi.org/10.24381/Cds.E2161bac>, 2019.
- Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzi, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global Prediction of Extreme Floods in Ungauged Watersheds, *Nature*, 627, 559–563, <https://doi.org/10.1038/s41586-024-07145-1>, 2024.
- 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.
- Rakhymbek, K., Mukanova, B., Bondarovich, A., Chernykh, D., Alzhanov, A., Nurekenov, D., Pavlenko, A., and Nugumanova, A.: LSTM-Based River Discharge Forecasting Using Spatially Gridded Input Data, *Data*, 10, <https://doi.org/10.3390/data10080122>, 2025.
- Rew, R., Hartnett, E., and Caron, J.: NetCDF-4: Software Implementing an Enhanced Data Model for the Geosciences, in: 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, vol. 6, 2006.
- Snow, A. D., Cochran, M., Miara, I., Hoese, D., den Bossche, J. V., Mayo, C., Lucas, G., Cochran, P., de Kloe, J., Karney, C., Shaw, J. J., Anh, T. Q., Filipe, Ouzounoudis, G., Couwenberg, B., Lostis, G., Dearing, J., Jurd, B., Gohlke, C., Schneck, C., McDonald, D., Taves, M., Itkin, M., May, R., Stewart, A. J., de Bittencourt, H., Little, B., Hugonnet, R., and Rahul, P. S.: Pyproj4/Pyproj: 3.7.2rc1, Zenodo [code], <https://doi.org/10.5281/zenodo.16817340>, 2025.
- Van Rossum, G. and Drake Jr, F. L.: Python 3 Reference Manual, Scotts Valley: CreateSpace, ISBN 978-1-4414-1269-0, 2009.
- Vischer, M. A., Otero, N., and Ma, J.: Spatially Resolved Meteorological and Ancillary Data in Central Europe for Rainfall Streamflow Modeling, *HydroShare* [data set], <https://doi.org/10.4211/hs.d7f2cbb587ab4a75ac7987854e8f62ca>, 2025a.
- Vischer, M. A., Otero, N., and Ma, J.: Spatially resolved rainfall streamflow modeling in central Europe, *Hydrol. Earth Syst. Sci.*, 29, 5233–5250, <https://doi.org/10.5194/hess-29-5233-2025>, 2025b.