# Data Descriptor Template

**Scope Guidelines**

**Data Descriptors** submitted to *Scientific Data* should provide detailed descriptions of valuable research datasets, including the methods used to collect the data and technical analyses supporting the quality of the measurements. Data Descriptors focus on helping others reuse data, rather than testing hypotheses, or presenting new interpretations, methods or in-depth analyses. Relevant datasets must be deposited in an appropriate public repository prior to Data Descriptor submission, and their completeness will be considered during editorial evaluation and peer review. The data must be made publicly available without restriction in the event that the Data Descriptor is accepted for publication (excepting reasonable controls related to human privacy issues or public safety).

### Title

*110 characters maximum, including spaces*

Titles should avoid the use of acronyms and abbreviations where possible. Colons and parentheses are not permitted.

Multiorder Hydrologic Position in Europe as a Set of Metrics in Support of Groundwater Mapping at Regional and National Scales

### Authors

Maximilian Nölscher1, Maximilian Nölscher2, Stefan Broda3

**Affiliations**

1. Federal Institute for Geosciences and Natural Resources (BGR), Berlin, Germany

2. independent researcher

3. Federal Institute for Geosciences and Natural Resources (BGR), Berlin, Germany

corresponding author(s): Maximilian Nölscher (max-n@posteo.de)

### Abstract

*170 words maximum*

The Abstract should succinctly describe the study, the assay(s) performed, the resulting data, and their reuse potential, but should not make any claims regarding new scientific findings. No references are allowed in this section.

This dataset (EU-MOHP v013.0.1) provides information on the multiorder hydrologic position of a geographic point within its respective river network or catchment. More precisely, it comprises the three measures “lateral position” as a relative measure of the position between the stream and the catchment boundary/ watershed, “divide stream distance” as an absolute distance measure that serves as a proxy for the position within the catchment and “stream distance” as an absolute measure of the distance to the nearest stream. These three measures were calculated for several hydrologic (stream) orders. Its spatial extent covers major parts of physiographical Europe and all of the 39 countries in European Economic Area (EEA39). Although there might be many potential use cases, this dataset serves predominantly as valuable input data for mapping tasks in the context of hydrogeology and subsurface characteristics in general.

### Background & Summary

*700 words maximum*

The Background & Summary should provide an overview of the study design, the assay(s) performed, and the data generated, including any background information needed to put this study in the context of previous work and the literature, and should reference literature as needed. The section should also briefly outline the broader goals that motivated collection of the data, as well as their potential reuse value. We also encourage authors to include a figure that provides a schematic overview of the study and assay(s) design.

[Background]

In recent years, data science tools such as machine learning are increasingly applied to and specifically developed for hydro(geo)logical challenges and research questions. In the field of hydrogeology, machine learning has been used successfully for groundwater level prediction and a variety of mapping tasks. Since machine learning models are traditionally based purely on data with no built-in knowledge of physical processes, it is important to provide as many variables (predictor variables/ explanatory variables/ features) as possible that have an impact on the target variable to potentially enable the machine learning algorithm to reproduce the result of the underlying process. For surface and near-surface processes, this criterion may be more or less satisfiable through the availability of remote sensing data, whereas for modelling subsurface processes such as in hydrogeology, this poses a serious challenge.

[Motivation/goals]

The key motivation of this dataset is to generate a set of features that function as a proxy for multiple geophysical characteristics of a hydrologic system and therefore complements commonly available datasets and tackles the above mentioned challenge.

This dataset is strongly inspired by [!!source: Beelitz et. al.] and adapts their ideas and methods to the “EU-Hydro - River Network Database” but – in contrast – with purely free open source software and a strong focus on reproducibility. [!!source: Beelitz et. al.] provides a comprehensive explanation of the motivation as well as a detailed discussion for further reading.

[context of previous work and the literature]

In their study, [!!source: Beelitz et. al.] also provide the results from case studies to proof that the multiorder hydrologic position is a valuable feature when mapping divers geophysical targets using machine learning. Its benefit to the performance of machine learning models has also been acknowledged by several other studies [!!source: Using Boosted Regression Tree Models to Predict Salinity in Mississippi Embayment Aquifers, Central United States; Machine Learning Predictions of pH in the Glacial Aquifer System, Northern USA; Machine-learning models to map pH and redox conditions in groundwater in a layered aquifer system, Northern Atlantic Coastal Plain, eastern USA; The relation of geogenic contaminants to groundwater age, aquifer hydrologic position, water type, and redox conditions in Atlantic and Gulf Coastal Plain aquifers, eastern and south-central USA].

[potential reuse]

* Feature for mapping hydrogeochemical parameters
* Feature for mapping hydraulic variables such as depth to groundwater
* Feature for groundwater level prediction
* Catchment classification/clustering

### Methods

The Methods should include detailed text describing any steps or procedures used in producing the data, including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). See the [detailed section in our submission guidelines](https://www.nature.com/sdata/publish/submission-guidelines#sec-5) for advice on writing a transparent and reproducible methods section. Related methods should be grouped under corresponding subheadings where possible, and methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study. Specific data outputs should be explicitly referenced via data citation (see Data Records and Citing Data, below).

Authors should cite previous descriptions of the methods under use, but ideally the method descriptions should be complete enough for others to understand and reproduce the methods and processing steps without referring to associated publications. There is no limit to the length of the Methods section.

[Disclaimer]

All processing and analysis was conducted with free open source software. All processing steps except for the download of the “EU-Hydro - River Network Database” (for practical reasons also referred to as river database) that was done manually are controlled and executed from within a targets pipeline in the programming language R [!!source]. Targets is an R package that provides a toolkit for reproducible workflows [!!source]. Spatial vector data such as the EU-Rivers are processed partly in R and a PostgreSQL (version 13) database with PostGIS (version 3.1.0) extension database for speed and memory reasons. For the same reason, all major raster calculations were conducted in a GRASS GIS (version 7.8.5-2) database. The database connections and all calculations in the databases are also controlled by this pipeline. For reaching a maximum of reproducibility, a docker container is provided to rerun all calculations easily. The R package renv is used for keeping track of the required R package versions and fits well to the combination with targets and docker to endure reproducibility.

[Methods main]

In the following, the description of the methods is oriented towards the structure of the targets pipeline to easily relate the text to the source code. All required steps to understand the workflow will be described, for further details we refer to the source code.

Step 1: Data Acquisition

The “EU-Hydro - River Network Database” was manually downloaded from https://land.copernicus.eu (for detailed link see references) as version v013. All download files extracted have approximately 14 GB. The river database is the only underlying data for the generation of the EU-MOHP dataset.

[Hardware]

The pipeline to generate the dataset was executed on a DELL PowerEdge C4140 Server with an Intel Xeon Gold 6240R CPU and 384 GB installed RAM. The installed operation system is Microsoft Windows Server 2019 Standard, version 10.0.17763 Build 17763.

[what is different to the Beelitz Paper and why]

* No pathleveId column
* Criterion to exclusively use free open source software

### Data Records

The Data Records section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats. Each external data record should be cited as described below. A data citation should also be placed in the subsection of the Methods containing the data-collection or analytical procedure(s) used to derive the corresponding record.

Tables should be used to support the data records, and should clearly indicate the samples and subjects (study inputs), their provenance, and the experimental manipulations performed on each. They should also specify the data output resulting from each data-collection or analytical step, should these form part of the archived record.

### Technical Validation

The Technical Validation section should present any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by figures and tables, as needed. *This is a required section*; authors must provide information to justify the reliability of their data.

Possible content **may include:**

* experiments that support or validate the data-collection procedure(s) (e.g. negative controls, or an analysis of standards to confirm measurement linearity)
* statistical analyses of experimental error and variation
* phenotypic or genotypic assessments of biological samples (e.g. confirming disease status, cell line identity, or the success of perturbations)
* general discussions of any procedures used to ensure reliable and unbiased data production, such as blinding and randomization, sample tracking systems, etc.
* any other information needed for assessment of technical rigour by the referees

Generally, this **should not include:**

* follow-up experiments aimed at testing or supporting an interpretation of the data
* statistical hypothesis testing (e.g. tests of statistical significance, identifying differentially expressed genes, trend analysis, etc.)
* exploratory computational analyses like clustering and annotation enrichment (e.g. GO analysis).

### Usage Notes

*This section is optional*

The Usage Notes should contain brief instructions to assist other researchers with reuse of the data. This may include discussion of software packages that are suitable for analysing the assay data files, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Please see our [code availability policy](http://www.nature.com/sdata/policies/editorial-and-publishing-policies#code-avail) for advice on supplying custom code alongside Data Descriptor manuscripts.

For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.

### Code Availability

For all studies using custom code in the generation or processing of datasets, a statement must be included under the subheading "Code availability", indicating whether and how the code can be accessed, including any restrictions to access. This section should also include information on the versions of any software used, if relevant, and any specific variables or parameters used to generate, test, or process the current dataset.

### Acknowledgements

The Acknowledgements should contain text acknowledging non-author contributors. Acknowledgements should be brief, and should not include thanks to anonymous referees and editors or effusive comments. Grant or contribution numbers may be acknowledged.

### Author contributions

Each author’s contribution to the work should be described briefly, on a separate line, in the Author Contributions section.

### Competing interests

A competing interests statement is required for all papers accepted by and published in *Scientific Data*. If there is no conflict of interest, a statement declaring this must still be included in the manuscript.

### Figures

Figure images should be provided as separate files and should be referred to using a consistent numbering scheme through the entire Data Descriptor. In most cases, a Data Descriptor should not contain more than three figures, but more may be allowed when needed. We discourage the inclusion of figures in the Supplementary Information – all key figures should be included here in the main Figure section.

For initial submissions, authors may choose to supply a single PDF with embedded figures.

Authors are encouraged to consider creating a figure that outlines the experimental workflow(s) used to generate and analyse the data output(s).

### Figure Legends

### Figure legends begin with a brief title sentence summarizing the purpose of the figure as a whole, and continue with a short description of what is shown in each panel and an explanation of any symbols used. Legends must total no more than 350 words, and may contain literature references. The first sentence of the legend will be used as the title for the figure. It should contain no references of any kind, including to specific figure panels, bibliographic citations or references to other figures or panels.

### Tables

Authors are encouraged to provide one or more tables that provide basic information on the main ‘inputs’ to the study (e.g. samples, participants, or information sources) and the main data outputs of the study; also see the additional information on providing metadata on page 6. Tables in the manuscript should generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an appropriate data repository.

Authors may provide tables within the Word document or as separate files (tab-delimited text or Excel files). Legends, where needed, should be included in the Word document. Generally, a Data Descriptor should have fewer than ten tables, but more may be allowed when needed. Tables may be of any size, but only tables that fit onto a single printed page will be included in the PDF version of the article (up to a maximum of three).

Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the article and will be made available in the online version only. Any such tables must be labelled in the text as ‘Online-only’ tables and numbered separately from the main table list e.g. ‘Table 1, Table 2, Online-only Table 1’ etc.

### References

Bibliographic information for any works cited in the above sections, using the standard Nature referencing style.

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference. See ‘Citing Data’ below for further details.

### Additional Formatting Information

**Referencing Figures, Tables, and other content**

**The Word document may reference Figures (e.g. Fig. 1), Tables (e.g. Table 1), online-only tables (e.g. Online-only Table 1) and Supplementary Information (e.g. Supplementary Table 1, or Supplementary File 2, etc.). When information from metadata documents must be referred to, it should also be included in the main manuscript as Tables, and formatted in a way that suits human readability. To refer to the ISA-Tab metadata records within the manuscript, use the phrase “see associated Metadata Record”.**

**Citation format**

All references should be numbered sequentially, first throughout the text, then in tables, followed by figures and, finally, boxes; that is, references that only appear in tables, figures or boxes should be last in the reference list. Only one publication is given for each number. Only papers that have been published or accepted by a named publication or recognized preprint server should be in the numbered list; preprints of accepted papers in the reference list should be submitted with the manuscript. Published conference abstracts, numbered patents, and archived code with an assigned DOI may be included in the reference list. Grant details and acknowledgments are not permitted as numbered references. Footnotes are not used.

Scientific Data uses standard Nature referencing style. All authors should be included in reference lists unless there are six or more, in which case only the first author should be given, followed by ‘et al.’. Authors should be listed last name first, followed by a comma and initials (followed by full stops, '.') of given names. Article titles should be in Roman text; only the first word of the title should have an initial capital and the title should be written exactly as it appears in the work cited, ending with a full stop. Book titles should be given in italics and all words in the title should have initial capitals. Journal names are italicized and abbreviated (with full stops) according to common usage. Volume numbers and the subsequent comma appear in bold. The full page range should be given where appropriate. See the examples below:

**Journal Article**:

1. Schott, D. H., Collins, R. N. & Bretscher, A. Secretory vesicle transport velocity in living cells depends on the myosin V lever arm length. *J. Cell Biol*. **156**, 35‐39 (2002).

**Book** ‐ Book titles should be given in italics and all words in the title should have initial capitals:

1. Hogan, B. *Manipulating The Mouse Embryo: A Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, 1994)

**Publicly available preprint:**

1. Babichev, S. A., Ries, J. & Lvovsky, A. I. Quantum scissors: teleportation of single-mode optical states by means of nonlocal single photon. Preprint at http://arXiv.org/quant-ph/0208066 (2002).

**Code:**

1. Gallotti, R. & Barthélemy, M. Source code for: The multilayer temporal network of public transport in Great Britain. *Figshare* https://dx.doi.org/10.6084/m9.figshare.1249862.v1 (2014).

**Online material** ‐ Stable documents hosted on the web may be cited in the main reference list, using the format below. Websites or dynamic web resources should be cited by embedding the URL in the main article text:

1. Manaster, J. Sloth squeak. *Scientific American Blog Network* http://blogs.scientificamerican.com/psi-vid/2014/04/09/sloth-squeak (2014).

**Technical or government report:**

1. Akutsu, T. *Total Heart Replacement Device.* Report No. NIH-NHLI-69 2185-4 (National Institutes of Health, 1974).

## Citing Data

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference.

An author list (formatted as above) and title for the dataset should be included in the data citation, and should reflect the author(s) and dataset title recorded at the repository. If author or title is not recorded by the repository, these should not be included in the data citation. The name of the data-hosting repository, URL to the dataset and year the data were made available are required for all data citations. For DOI-based (e.g. figshare or Dryad) repositories the DOI URL should be used. For repositories using accessions (e.g. SRA or GEO) an [identifiers.org](https://identifiers.org/) URL should be used where available. For first submissions, authors may choose to include just the accession number. Scientific Data staff will provide further guidance after peer-review. Please refer to the following examples of data citation for guidance:

1. Zhang, Q-L., Chen, J-Y., Lin, L-B., Wang, F., Guo, J., Deng, X-Y. Characterization of ladybird Henosepilachna vigintioctopunctata transcriptomes across various life stages. figshare <https://doi.org/10.6084/m9.figshare.c.4064768.v3> (2018).
2. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRP121625> (2017).
3. Barbosa, P., Usie, A. and Ramos, A. M. Quercus suber isolate HL8, whole genome shotgun sequencing project. GenBank<http://identifiers.org/ncbi/insdc:PKMF00000000> (2018).
4. DNA Data Bank of Japan <http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA004814> (2016).

**Depositing your data to an appropriate repository**

Your *Scientific Data* manuscript will not be sent to review unless the dataset(s) described therein have been deposited in an appropriate public repository ([please see our list of recommended repositories](http://www.nature.com/sdata/policies/repositories)). Should a specific repository not be available for your field or data-type, or should the repository of your choice not permit confidential peer-review, you may upload your data to one of our [recommended generalist repositories](https://www.nature.com/sdata/policies/repositories#general). Integrated submission systems are available for both figshare and Dryad.