Plan Overview

A Data Management Plan created using DMPonline

Title: Hydrological Modelling of Groundwater-Surface Water Interactions and Climate Change

Creator: Marios Karampasis

Principal Investigator: Marios Karampasis

Data Manager: Marios Karampasis

Affiliation: Utrecht University

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: UU Data Management Plan (DMP)

ORCID iD: 0009-0009-2149-9422

Project abstract:

This project is WP3 of the WaterScape project (WaterScape | Universiteit Utrecht | Homepage). The goal of the project is to enhance existing hydro(geo)logical models (provided by Deltares), to model the impact of climate change on 3 Dutch regions (living labs).

Potential modifications to the hydrological system will be evaluated via modelling scenarios to provide solutions to the impact of climate change for the stakeholders involved.

ID: 177431

Start date: 15-07-2024

End date: 14-07-2028

Last modified: 07-05-2025

Hydrological Modelling of Groundwater-Surface Water Interactions and Climate Change

Data Collection

1.1 Will you re-use existing data ?

If yes: explain which existing data you will re-use and under which terms of use.

- Yes, I will re-use existing data

The hydrological models used in this project (by the primary investigator - PhD candidate) have been developed and are managed by Deltares. The models are based on parameter sets including:
a. geospatial data - mostly square grids, e.g. geological layer elevation.
b. parameters representing the whole model area, e.g. runoff factor for residential areas.
c. geospatial point timeseries (TS), e.g. borehole abstraction records.
d. geospatial grid TS, e.g. precipitation records.
e. configuration parameters, specifying in which way the data will be used by the software (iMOD).

The data was provided by Deltares for research purposes. They are based on open datasets from governmental organizations (e.g., geological surveys, meteorological institutes) and are used in accordance with their respective open data licenses.

The volumes described below apply to only 1 out of 3 model used for this project, as the investigation on the other two study areas will only begin at a later stage. The data volumes for the remaining models are expected to be comparable to the first model (North Brabant model).

1.2 Describe your data.

Fill the table below with a brief description of the data, including the type, format and volume.

#	Data Description	Data Type	Format	Total Volume
1	Geological layer bottom/top elevation	geospatial square grids	.idf	448 MB
2	Unsaturated zone parameters	categorical grids (e.g. land uses) tabular (e.g. model configuration table) global parameters (e.g. urban area infiltration factor)	.inp .idf	828 MB
3	Meteorological TS (precipitation, potential evapotranspiration)	geospatial grid TS	.asc	33 GB
4	Constant head outputs from larger models (used as boundary conditions)	geospatial grids	.idf	34 GB
5	Drain elevation and conductance	geospatial grids	.idf	1 GB
6	Horizontal flow barriers	geospatial lines	.gen	20 KB
7	Geological formation hydraulic conductivity	geospatial grids	.idf	245 MB
8	Borehole hydraulic head observations	geospatial point TS	.ipf .txt	205 MB
9	Project files (configuration files)	text files	.prj	400 KB
10	River parameters (bed bottom, stage, conductance)	geospatial grids	.idf	870MB
11	Starting hydraulic heads	geospatial grids	.idf	123 MB
12	Borehole abstraction rates	geospatial point TS	.ipf	10 MB

Data Documentation

2.1 Describe the documentation and metadata that you will use to to make your data reproducible and interoperable.

Describe which files you will provide, along with a brief description of the information they will contain, to make your data reproducible and interoperable. Describe the information that you will provide to make the data items in questions 2.1 reusable and interoperable. If using a specific metadata standard, please mention this below.

Geospatial files (.idf) contain metadata internally in the file header (similarly to TIFF or NetCDF files).

Geospatial TS files (.asc, .ipf) contain metadata as text (first few lines) and are named based on the date they represent, e.g. a precipitation file: P_1991010_NBr1.asc (there is an acronyms.xlsx file where abbreviations are defined (P: Precipitation), NBr1 is the model number (each simulation gets one based on the model alias (NBr: North Brabant (defined in the model run log (WS_RunLog.xlsx)), and the Simulation number (1)).

The model RunLog contains information about each model simulation (Sim): which Sim it uses as a baseline, which parameters were modified, when it was run etc. Each Sim is linked to a snakemake (.smk) file, which sets the workflow used to create the output (ensuring reproducibility).

Configuration files (.prj, .inp) contain information in very specific formats, as described by the corresponding software's (iMOD, Modflow, MetaSWAP) documentation, which are stored in the same folder.

Read_me (.md or .txt) files stored in the project folder contain the aforementioned information, along with other details, useful for anyone involved with the project.

2.2 Describe the folder structure you will provide to make your data reproducible and interoperable.

Describe the folder structure, naming conventions and/or version control you will use for this project.

Copied from my README.md file:

Introduction
This project folder is used for the purposes of WaterScape Regional Groundwater modelling.
From this point on, abbreviations/acronyms defined in .\Mng\Acronyms.xlsx are used (to save time and space).
Initially, this folder is being set up in my (Marios Karampasis) OneDrive. The project software requires windows to run, so unfortunately it's only possible to work on this project from a Win machine. To avoid potential errors caused by OneDrive's default name (includes a hyphen and spaces), make a permanent symbolic link on your C: drive using the following command in your CMD:
mklink /D C:\OD "C:\Users\<User>\OneDrive - Universiteit Utrecht"
(change the OneDrive directory to whatever it needs to be)

The symbolic link can be deleted using:
rmdir C:\OD

All files you need to run the models should be available within this project folder. The only thing you need to install yourself is the python environment - use the guide below.

Rules and Tips
- .\Mng\Acronyms.xlsx contains acronyms and rules used throughout the project. This is not just an acronym archive. It also explains the naming convention used throughout this project, so you need to be familiar with its principles when working on with this folder/project. New acronyms should be registerred there. Make sure you read the instructions sheet/tab to understand the acronym system, and follow the system's rules when adding new shortcuts.
- The folder is version controlled through GitHub (smaller files, e.g. code) (and DVC). You can use "git ls-files" to print all Git tracked files, and "dvc list . --dvc-only --recursive" (warning, it's very slow), to print the files tracked with each of the two methods. Make sure you push at frequent intervals.
- For good data management, it's advised to include meta-data in data folder that are not self-explanatory. E.g. a read-me file in a folder with IDFs (IDFs may contain spatial data, units etc. but oftentimes their origin/method of production, which can be very useful, is missing from the meta-data).

Folder structure/description
Files that are specific to one of the models will be contained withing the folder of that model. The rest of the folders in this directory should contain files that are (or will be) used by/for multiple models.

Below is a brief description of what is contained within each of the main folders of this directory.
- code: Contains scripts and code. Sub-folders grouped by function/purpose.
- data: Contains data not specific to one model (e.g. KNMI climate TS).
- Mng: Contains files used for managing the project. WS_RunLog.xlsx is used to keep track of runs. It reads log.csv, where Sim info is recorded as the Sim is being executed.
- models: Contains a sub-folder for each model used in the project. All data/files related to just one of the models belong inside those sub-folders. If no metadata about the file source is provided, asume the source is the Deltares P: drive (e.g. for NBr: under p:\archivedprojects\11209224-sponswerkingnl\ or p:\archivedprojects\11206534-002-imod-brabantsedelta\)
the models folder structure is described in more detail below because it is complex and critical for the project.
- other: Files relevant to the project that don't belong in any of the other categories/folders.
- software: Contains modelling software
- SS: Superseded - anything not relevant anymore. (although superseded files can be found in other sub-folders too.

models
All model sub-folders contain the same Fo Str for consistency. Files in those folders are only relevant to this Mdl. The Fo Str is described below:
- code: Contains code specific to each Mdl. e.g. Mdl_Prep contains the .bat & .ini file to prepare a Mdl run.
- doc: self explanatory
- In: model inputs, organized by iMOD package/module (organized by type). Only contains files that are used directly in the model - i.e. no raw data etc.
- MM: Contains 3 types of elements:
- general information layers, e.g. rivers, background map etc. - Those elements are used in the MM regardless of run.
- In files converted to GIS layers for review and visualization. Most inputs remain the same as the B for a run, but some have to be re-referenced on each Sim. This is done by changing the data-source of the layer (programmatically).
- PoP: contains a model map for each Sim. PoPed output is unique to each Sim. Thus, the B MM gets copied and the output layers, which are relatively referenced, automatically get linked to the new Sim's PoPed Out.
- PrP: Pre-Processing. Can include raw or intermediate data, or even scripts/JupNotes to create the In files.
- Sim: Simulation folders. 1 for each Sim. Unfortunately, the way iMOD is designed, Mdl Output needs to be saved here. Organized by Sim.

Python Env installation guide
Follow the steps below, to install the WS python Env. Then use it whenever you want to run any command related to this project. This way all dependencies will be satisfied.
1. Each Env version is linked to a MdlN (more about MdlNs in the RunLog (./Mng/WS_RunLog.xlsx). Replace <MdlN> below with the Env version you want to install. Differenes between versions are miniscule, but it's advised to stick to the Env version that corresponds to your Sim. Not all Sims have a unique Env, because it's not always necessary to make ammendments to the previous Env. In this case use the latest Env before the Sim you want to execute.
mamba env create -f Env_<MdlN>.yml

2. Install this project's python library. this assumes that you've made a symbolic link, as described above. Otherwise replace the path.
pip install -e C:\OD\WS_Mdl\code (or pip install -e C:\OD\WS_Mdl\code --use-pep517)

Terminal tools

There is a list of terminal tools that facilitate common tasks for this project. Those are listed in C:/OD/WS_Mdl/code/setup.py, with a brief description.
To add another terminal command, you need to add it to the setup file (similar to the other commands), and make a script. Then you need to run step 2 from the python Env installation guide above.
It's also possible to run python function from C:\OD\WS_Mdl\code\WS_Mdl\WS_Mdl.py via "WS_Mdl <function> <arg1> <arg2> ...".

Data Storage

3.1. Select the storage solution(s) where you will store and back-up your data.

Select the locations where your data will be stored. You may select more than one. Please describe the storage solution and the backup strategy of your storage solution if it does not appear in the list below.

- OneDrive for Business
- SURF Research Drive
- YoDa

SURF Research Drive will be used for immediate storage of model output. Raw outputs will be post processed and uploaded to YoDa long term (with metadata). Raw output will be deleted after it's been post processed, as it can aways be recreated by running the same snakemake file (the model is deterministic). Only part of the raw data will be kept, for cross-checking purposes.

Needless to say, all model input and other files needed to recreate the output will also be stored on YoDa.

OneDrive is used to back-up any files related to my work (i.e. not just the modelling folder). The model folder is stored there, but just temporarily, as OneDrive causes errors sometimes - when the required input files aren't immediately available.

Data Privacy and Security

4.1 Will you be collecting or using personal data ?

Personal data is any data which, alone or in combination with other information, can identify a living person. Such data must abide by the GDPR and requires additional safeguards and documentation to be processed lawfully.

- No, I will not collect and/or use personal data

4.2 How will ownership and intellectual property rights of the data be managed?

Describe who controls access to the data and who determines what is done to the data.

The Principal Investigator (PI) will determine and manage access to the research data. Intellectual property rights will remain with Utrecht University.

Throughout the project, all members of the research group, including students, and consortium members of the WaterScape project will have access to the data. It's possible for external researchers to gain access, but it must be explicitly granted by the PI (in agreement with the consortium members).

Data Selection, Preservation & Sharing

5.1 Describe the data you will be preserving and the storage solution where it will be preserved?

Describe which data will be preserved under long-term storage. You may refer back to the data described in question 1.2 to specify which data will be preserved. Explain where you will preserve your data, and how procedures are applied to ensure the survival of the data for the long term.

As described in section 3.1, the simulations will be run on SURF Research Cloud's drive(s). All the input files (necessary to create the model output) will be stored there to reduce run-times. The raw output will be produced there, then it'll be post processed, and mostly deleted.

YoDa will be used for long term storage of model inputs and post processed outputs (which are much smaller than raw outputs). A small part of the model outputs will also be stored to allow cross-checking in the future.

All inputs are also stored on the primary investigator's laptop and they're backed up via OneDrive. This covers the 3+2+1 rule. >=3 copies: local, OneDrive, Yoda. >=2 media: local storage, cloud (2 (OneDrive, YoDa). >=1 copy off site - cloud (2).

Input data have been described already in a previous chapter. Post Processed outputs include: geospatial TS files, statistics, plots, GIS layers, etc.

5.2 Describe the data you will be sharing and the repository where it will be shared?

Describe which data you will be sharing. Select where you will make your data findable and available to others. If selecting "Other" please specify below which repository and provide a URL.

Please also write below if you will apply any conditions to the re-use of your data. (i.e. Creative commons license or Data Transfer Agreement).

- YoDA

Data is version controlled by GitHub + DVC. To gain access locally, other researchers can clone the repository locally, then download the large input files (which aren't available on GitHub) from Yoda.

The data can be re-used within the scope of the project. For use outside the project, explicit permission needs to be provided by the PI.

5.3 Are specialized, uncommon or expensive software, tools or facilities required to use

the data?

Please list any specialized, uncommon or expensive software, tools or facilities that are absolutely required to obtain, use or handle your data, if any.

No, all software is open source and will be uploaded to YoDa along with the input.
The users are required to read the terms of use of each software and agree to them.

Data Management Costs and Resources

6.1 What are the foreseeable research data management costs and how do you expect to cover them ?

Please specify the known and expected costs involved in managing, storing and sharing your data. Also explain how you plan to cover these costs.

There are some potential costs related to data storage (mainly long-term storage). A limited amount of data can be stored for free on YoDa and OneDrive. If the data exceeds that amount, the incurred costs will be covered from the project's budget.
E.g. 10 TB for 10 years * 12 months * 4 €/TB/month = 4800 €

6.2 Who will be responsible for data management?

Please specify who is responsible for updating the DMP and ensuring it is being followed accordingly.

The PI is responsible for the updating and ensuring the DMP.

6.3 State if you contacted an RDM consultant from Utrecht University to help you fill out your DMP.

Please list their name and date of contact.
This is mandatory for NWO grants.

The PI contacted 3 Geosciences Data Management members:
Aristotelis Kandylas and Garrett Speed provided consultation regarding choosing the best platform fo the project's needs.
Vincent Brunst was provided access to YoDa.
There have also been consultation with Jelle Treep, from the Research Engineering team regarding usage of SURF Research Cloud services for HPC and storage purposes.