Data Management Plan

Name: Lucie C. Vermeulen MSc
Group: Environmental Systems Analysis Group (ESA), Wageningen University
Supervisor: Dr. ir. Nynke Hofstra (ESA)
Promotors: Prof. dr. Carolien Kroeze (ESA & Open University), Prof. dr. Gertjan Medema (TU Delft & KWR Watercycle research institute)

Short description of research

Title: A Spatially Explicit Modelling Approach to Estimate Waterborne Pathogen Concentrations in the Surface Waters of the World.

I plan to develop a spatially explicit process-based global scale waterborne pathogen model using existing datasets on population density, land use, livestock and hydrology, among others. The form of this model will be based on existing global nutrient models e.g. (Bouwman et al. 2006) and the first exploration prepared Hofstra et al. (Hofstra et al. 2013). I will focus on *E. coli* and *Cryptosporidium.* I hope to obtain existing data sets on concentrations of these pathogens in surface water for validation.

Four focus areas of the PhD research, each expected to result in a paper:

1. Comparing existing modelling efforts in global nutrient models and waterborne pathogen models, identifying difficulties and opportunities
    - Methodology: literature study
2. Designing and validating a global waterborne pathogen model
    - Methodology:
    1) drafting a conceptual model, selection and adaptation of existing model components that can serve as basis,
    2) gathering data to determine model parameter values (e.g. pathogen prevalence rates, sewage treatment levels), gathering gridded input data sets (e.g. land use data, population and livestock density, climate data, hydrological data), checking these for errors and make them into usable format,
    3) model coding in R, testing the model
    4) model validation, by using existing data sets of waterborne pathogen measurements, sensitivity analysis
3. Model application
    - Methodology: scenario analysis using the Millennium Ecosystem Assessment scenarios, as has been done by global nutrient modelling studies
4. Extending the scope
    - Methodology: literature analysis into how this model could be applied in for example risk assessment studies

Data management roles

I am responsible for collecting and managing the data used in my research, in consultation with my supervisor. At the end of my PhD project my data will be made publically available online if possible, at least my supervisor will receive a full copy of the data.

Types and amount of research data

I will collect existing publically available datasets, as well as not publically available pathogen measurements.
I will produce a model, and with this produce maps and estimates of waterborne pathogen concentrations in surface waters worldwide.
I estimate to have up to 20 GB of research data. This estimate is based on my MSc thesis which is in the same subject area, which used around 1,5 GB of data.

Type of research data	**Specification**	**Software choices**	**File extension**
Model parameter values	I will need to gather information to use as model parameter values, for example on pathogen removal rates for different types of sewage treatment, or the die-off under specific environmental conditions. I plan to make an overview file of different parameter values found in literature or from other sources.	Excel	.csv
Model input data - gridded	Existing datasets on for example climate and hydrology	Depending on how they are available. Runoff data are .grd files, for example, most others I don’t know yet	.grd and others
Model input data – country data	Existing datasets on for example population and livestock density, land use	Depending on how they are delivered, perhaps a spread sheet format. I may all convert these to csv files	.csv
Model input data - metadata	For all input data, one document containing all metadata will be created, specifying at least source, time period, region, measurement method, type of data, unit of measurement, access rights, date downloaded. All data will be checked for consistency, and any changes made to input data will be documented.	Excel	.csv
Data for model validation - waterborne pathogen measurements	Various mostly not yet identified sources: different institutes who perform waterborne pathogen measurements	Depending on how they are delivered, probably a spread sheet format. I may all convert these to csv files	.csv
Data for model validation - metadata	For all data for validation, one document containing all metadata will be created, specifying at least source, time period, region, measurement method, type of data, unit of measurement, access rights, date downloaded. All data will be checked for consistency, and any changes made to input data will be documented.	Excel	.csv
Original model scripts	Created by me	R	.R
Adapted model scripts	Created by me	R	.R
Model output - tables	Created by me	R	.csv
Model output - tests	Created by me	R	.csv / .txt
Model output - graphs	Created by me	R	.tiff / .eps
Model output - maps	Created by me	R	undetermined
Manuscripts	Created by me: manuscripts of papers	Word	.docx
Other	Administrative documents, coursework-related documents	Adobe, Word, Excel	.pdf /.docx / .xlsx

Sharing and ownership

I expect that others may want to use the model and its outcomes, and I intend to make these publically available.

Most, if not all, of the model input data are publically available.

Some pathogen measurement datasets I may use for validation probably are not. These could be made publically available in a combined database we are considering setting up. The specific requirements and set-up of such a database will need to be further delineated by me, my supervisor, and the owners of the respective datasets.

I have no specific funder or external party requirements regarding my data.

There are no privacy or security issues.

Documentation and metadata

During my research: When I start developing the model, different model versions will all be kept. With any publication of data, model version will be specified. For all input data, one document containing all metadata will be created, specifying at least source, time period, region, measurement method, type of data, unit of measurement, access rights, date downloaded. All data will be checked for consistency, and any changes made to input data will be documented.

For long term storage: For any publication of a model version or model output, the meta-information described above will be made available.

For data sharing: With any sharing of data, whether publically downloadable or upon request, the meta-information described above will be made available.

Short term storage

Format: The model will be in R, and the script can also be stored / made available as text file. Output data will be maps (some graphical format, jpg or tiff) and numerical data sets (could be as csv file). Input data for the model are xls or csv files, as well as gridded data in some raster type of file format, I do not know yet exactly what. Manuscripts will be written using current version of Word. Manuscripts intended for long-term storage can additionally be saved as txt file.

Current physical storage: The amount of data I have means that my M drive is not sufficient, and I am working on the local hard disk of my desktop now. I do weekly backups to a WUR network drive.

Long term storage

I intend to store my data long-term after the conclusion of my PhD project.

Intended way of storage: not yet decided. My chair group is currently designing a policy on this.

I consider perhaps using DANS (http://www.dans.knaw.nl/content/data-archief/data-deponeren), acknowledging that there are preferred storage formats, such as .pdf, .csv and .txt.

System for directory- and file-naming and version control

Manuscript versions and model version are numbered by using an extension _v1 etc. Old versions are kept in a folder titled Archive. Text documents will be named starting with the date in the form 20130416 and then the title. Model scripts will specify the type.

Data file names will start with an abbreviation of the type of data, then further specifics such as location and time, all separated by _. E.g. a gridded livestock density file of Europe in 2000 could be named liv_den_Europe_2000.csv

Proposed folder structure of PhD Lucie Vermeulen

Papers

- PDFs
- Paper MSc thesis 2012
- Paper 1
- Paper 2
- Paper 3
- Paper 4

Data

- Gridded
  - Hydrology
  - Point sources
  - Diffuse sources
  - Climate
  - Geophysical
- Country
  - Land use
  - Population
  - Livestock
  - Sewage treatment
- Pathogen measurements

Model

- Model scripts
- Model input
- Model output
  - Graphs
  - Tables
  - Maps
  - Tests

Administration

- Meeting agenda's and notes
- Financial
- Planning
- PhD proposal

SENSE

- General information
- A1 course
- Data Management course
- PhD Assessment
- …

Conferences

- Health Related Water Microbiology 2013
- …

Mischellaneous

- We Day
- IPCC review
- …

Archive