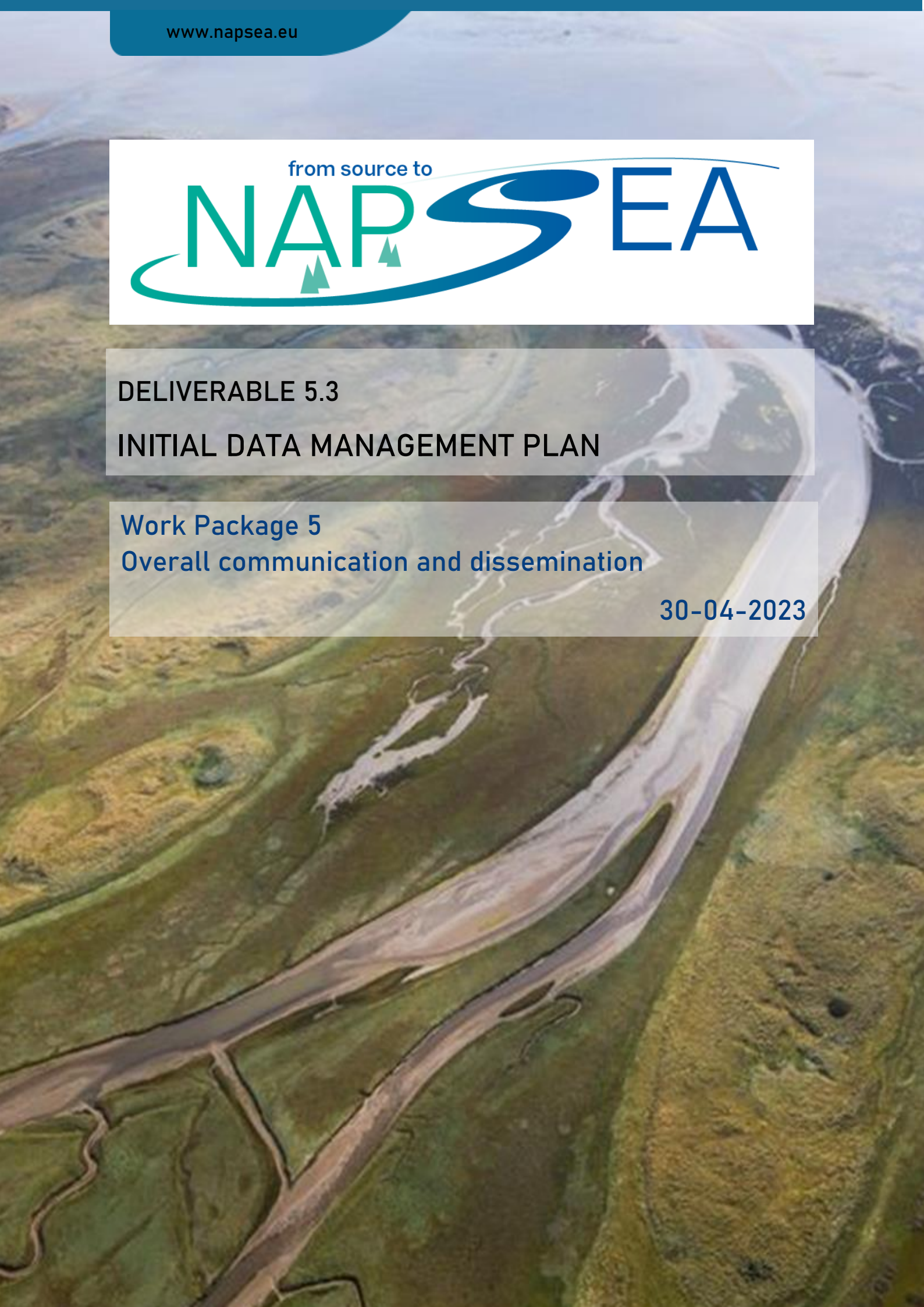

from source to NAPSEA

DELIVERABLE 5.3

INITIAL DATA MANAGEMENT PLAN

Work Package 5
Overall communication and dissemination

30-04-2023

Grant Agreement number	**101060418**
Project title	**NAPSEA: the effectiveness of Nitrogen And Phosphorus load reduction measures from Source to sEA, considering the effects of climate change**
Project DOI	
Deliverable title	Initial Data Management Plan
Deliverable number	D5.3
Deliverable version	Version 1.0
Contractual date of delivery	31-03-2023
Actual date of delivery	30-04-2023
Document status	Concept (of initial version)
Document version	Concept
Online access	Yes
Diffusion	Public
Nature of deliverable	Report
Work Package	WP5: Overall communication and dissemination
Partner responsible	Deltares
Contributing Partners	Deltares
Author(s)	Van Der Heijden, L.H.
Editor	Van Der Heijden, L.H.
Approved by	Blauw, A.
Project Officer	Blanca Saez-Lacava/Christel Millet
Abstract	The DMP is a document that outlines how data will be handled during and after the project, including what data will be collected, processed, and shared, as well as how it will be curated and preserved. The NAPSEA project will work with existing and new datasets, generate various types of data, and make the data findable, accessible, interoperable, and re-usable (FAIR data principle) through published deliverables. An updated version of the DMP will be delivered later in the project to incorporate lessons learned and barriers overcome.
Keywords	Data management; FAIR; Model output;

Contents

1. ACRONYMES

DG	Directorates-General
DMP	Data Management Plan
EC	European Commission
EU	European Union
FAIR	Findable; Accessible; Interoperable; Re-useable
HE	Horizon Europe
OA	Open Access
WP	Work package

2. EXECUTIVE SUMMARY

This deliverable will provide the guidelines and the first version of the Data Management Plan (DMP) for the NAPSEA project. Good data management is important in ensuring successful knowledge discovery and innovation, as well as subsequent data and knowledge integration and reuse by the community after the data publication process. The European Commission (EC) has provided guidelines for the data management plan that should include information on the handling of research data during and after the project, what data will be collected, processed and/or generated, which methodology and standards will be applied, whether data will be shared/made open access, and how data will be curated and preserved (including after the end of the project).

The NAPSEA project's DMP will serve as a working document that can be updated throughout the project, following the general template recommended by the EC HE guidelines that includes a data summary, specifying data collection objectives, data types, size, storage, and utility, as well as a description of FAIR data use, allocation of resources, data security, and ethical aspects.

The NAPSEA project will work with existing datasets of German and Dutch origin and will generate new data, ranging from questionnaires to modelled data of nitrogen and phosphorus. The data size and storage will vary depending on the data type, with time series data and gridded model data. The data will be primarily used internally for monitoring, operation, and development of a generic framework, and secondarily by a larger group consisting of researchers, decision-makers, national and international governments, NGOs, and companies. The model output generated throughout the NAPSEA project will be made findable, accessible, interoperable, and re-usable (FAIR data principle) through published deliverables that describe the model output, provide underlying data in the form of NetCDF or other formats, and include metadata that describe the data used and generated.

An updated version of the DMP will be delivered later in the NAPSEA project in order to incorporate the implementation process, lessons learned, and barriers overcome in data management.

3. INTRODUCTION

As mentioned in Wilkinson et al. (2016): “Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process”. The data management plan (DMP) is an important component in successfully managing data. They describe the data management life cycle, including discovery, collection, processing, generation and organization. Managing data and setting standards from the start will increase the value of the data that will be collected, processed and generated. The European Commission (EC) states that, under guidelines for the Horizon Europe (HE) projects, a good DMP should include information on:

- The handling of research data during and after the end of the project;
- What data will be collected, processed and/or generated;
- Which methodology and standards will be applied;
- Whether data will be shared/made open access and;
- How data will be curated and preserved (including after the end of the project).

This deliverable serves as a DMP that will help and guide all project partners to follow common standards. This initial DMP should guide partners in data management from the beginning of the project. However, this DMP will be a working document that can be updated throughout the project. Our DMP follows the general template recommended by the EC HE guidelines that includes: a data summary, specifying data collection objectives, data types, size, storage and utility; a description of FAIR data use, with a description of how we keep data findable, accessible, interoperable and re-usable according to the FAIR data policy (Wilkinson et al., 2016). Furthermore, chapters about allocation of resources, data security and ethical aspects will be presented. As this document will be updated throughout the project, a chapter NAPSEA Datasets is added with a preliminary outline of how all datasets should be described in the DMP.

4. DATA SUMMARY

4.1 Existing data

The NAPSEA project will work with existing datasets of German and Dutch origin. An overview table of the used existing data is given (see Deliverable 3.1 for full description).

Data type	Variable	Resolution	Period	Source
Model input data				
Meteorological data	Precipitation Temperature (mean, min, max)	Daily (1km)	1950-2020	BOKU: High-resolution (1 km) daily climate data (precipitation, minimum and maximum temperatures) (Cruz-Alonso et al., (2023)). https://doi.org/10.1016/j.envsoft.2023.105627
Hydrological data	Predicted discharge	Daily at 5x5 km	1950-2020, Scenarios until 2100	Predicted discharge using the mHM hydrological model results https://doi:10.1029/2008WR007327 https://doi:10.1029/2012WR012195
	Measured discharge	Daily at gauging stations	1950-2020	https://portal.grdc.bafg.de/
Soil data	Soil type			European Soil Database https://esdac.jrc.ec.europa.eu/content/european-soil-database-v20-vector-and-attribute-data
Agricultural data, Nitrogen inputs	- Nitrogen surplus including diffuse sources from non-agricultural land, - Application of fertilizer/manure, - Dates of farming practices.	Annual at 10 x 10 km	1850-2019	Batool et al. (2022) https://doi.org/10.1038/s41597-022-01693-
Land use depended diffuse phosphorous inputs	Dissolved phosphorus inputs	Average values	-	Yang et al. (2021) https://doi.org/10.1016/j.watres.2021.116887
Nutrient point sources	Total N and total P discharge from point sources	Only WWTP plants with population equivalents (PE) > 2000 are considered	average	In Germany Büttner et al. (2022) https://doi.org/10.1016/j.watres.2022.118382 EU scale: EEA (2020) UWWTD database
Morphological data	Digital Elevation Model	Resampled to (100 m)	-	SRTM (https://earthexplorer.usgs.gov/) A terrain elevation model was obtained from the Shuttle Radar Topography Mission (SRTM) sensor.
	Land use/ land cover		-	CORINE https://land.copernicus.eu/pan-european/corine-land-cover Land-cover data were derived from CORINE Land Cover 10 ha (https://gdz.bkg.bund.de/index.php/default/open-data.html, last accessed 20 Dec. 2022). These datasets were resampled to a spatial resolution of 100 m × 100 m for model simulations.
Model validation data				

Measured Nitrogen and phosphorus concentra-tions, surface water	**Elbe and Rhine** Nitrogen and Phosphorous species concentration		1968-2020	**Elbe and Rhine** Ebeling et al. (2022) https://doi.org/10.5194/essd-14-3715-2022 Virro et al. (2021) https://doi.org/10.5281/zenodo.5097436
	Hunze TN, TP and other water quality parameters (such as Chloride, dissolved oxygen, pH, NO_3, PO_4 etc)	Weekly/ monthly	2000-2023	**Hunze** Biweekly meaured concentrations data at nine gauging stations from upstream to downstream area of the Hunze catchment was collected from the responsible authorities.
Measured nitrogen concentration, groundwater	Nitrogen species concentration	bi-annually to annualy	1990-2017	EEA database https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-icm-1

The NAPSEA project will also generate new data, both from questionaries and from model simulations.

4.2 Types of new data

The NAPSEA project works with different types of data, ranging from questionaries related data to modelled data of nitrogen and phosphorus. Based on our current assessment the following data types are at least involved:

- Social data
  - Data from questionaries
- Modelled data
- Indicators / Indexes, related to safe ecological boundaries

4.3 Data size and storage

The data storage and sharing solution of choice should handle a variety of data and data formats that will require some capacity of local storage. There are two main types of data to store:

- Time series, data in a single point for a single property over time:
  - Point Time Series – Fixed location (x, y, z) varying time, e.g. weather station;
- Gridded data, data on a structure or unstructured grid (1D, 2D or 3D) that varies in time (e.g. data from model results):
  - Raw data (usually in NetCDF, GeoTIFF or HDF format);
  - Processed data (usually as an image).

The size of expected data varies largely between the different formats, for time series this is in the order of several megabytes, whereas for gridded model data this can be in the order of several gigabytes or terabytes.

4.4 Data Utility

Data will be primarily used internally for monitoring, operation and development of a generic framework. Mainly this data will be used by the pilot operators and work packages that require data analysis and processing.

Secondary, data can be useful to a larger group consisting of, but not limited to researchers, decision-makers, national and international governments, NGO's and companies.

During the project the data will be examined for external use and this document will be updated accordingly that will be presented as the final data management plan in deliverable 5.4.

5. FAIR DATA

Merely existing data is used on which the FAIR data principle is already applied. In 6. Other research outputs more information is provided on the FAIR data principle for other research outputs (e.g., model output).

6. OTHER RESEARCH OUTPUTS

An example of other research outputs that are generated (and potentially re-used) throughout the NAPSEA project is model output. Several model simulations will be conducted throughout the course of the NAPSEA project. The data management of these newly generated model consists of a similar principle as the FAIR data principle.

6.1 Findable model output

The model output will be made findable via the published deliverables, both in the form of a map (figure) as well as in the form of underlying data for such a map (e.g., NetCDF of post-processed model-output).

6.2 Accessible model output

Data will be made openly available via the publication of delivered reports connected to certain work packages. For example, in deliverable D3.2 calibrated and validated model will be described. The output of these models will be presented in a report and the underlying data of these maps/figure will be provided along with such a report.

6.3 Interoperable model output

Interoperable model output is guaranteed by well describing the data used and generated in several products of NAPSEA. The model output will in a NetCDF format which is interoperable. NetCDF files include metadata as an integral part of the datafile, ensuring that the dimensions remain consistent across all variables. Additionally, NetCDF files contain global attributes such as provenance links, NetCDF version information, and version history. As a result, NetCDF provides a solution to many data sharing challenges and facilitates the delivery of interoperable data from the outset. NetCDF is widely utilized in hydrological studies, with many models utilizing NetCDF as both input and output files, resulting in the sharing of accurate metadata.

6.4 Re-useable model output

Documentation of the software and methods used (information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.) will be stated in the delivered product. Also, the data lying underneath such a map (thus post-processed model output) will be provided along with the deliverable.

7. ALLOCATION OF RESOURCES

Since no new data is expected to come in via measurements, no cost will be made for producing, storing and managing data. Existing data is already stored and managed. Data coming from model output will be stored and made FAIR via the delivered products (e.g., reports and models).

8. DATA SECURITY

The security of research data is of utmost importance to ensure that the data is not lost, stolen, or accessed by unauthorized parties. Therefore, several measures will be taken to secure the data throughout the project.

8.1. Privacy

Any sensitive data, such as personally identifiable information (PII), will be handled with extra care. Access to this data will be restricted to authorized personnel only, and the data will be encrypted at rest and in transit. Data that is no longer needed will be securely deleted using appropriate methods.

8.2. Data Security

Secure Storage

All research data will be stored in a secure and backed-up location. The data will be stored on a password-protected network drive with restricted access, and the server will be backed up regularly to ensure that no data is lost in the event of a system failure. data security and storage, at this stage a mix of storage technologies will be explored (cloud, no cloud, SQL, NoSQL, etc.) in order to achieve the most effective solution. The most obvious solution in the future is to move the storage to the cloud but at this stage this solution may not be the most cost effective and other alternatives may need to be considered. In any case the data storage solution will ensure that the new paradigm of cloud computing is compatible with a future deployment via Data and Information Access Service (DIAS). Apart from this, several existing marine data infrastructure platforms will be

considered that have a solid data security. For example, SeaDataNet is including the important issues of trust which are addressed in data-based research: security, confidentiality, ownership, assured provenance, authenticity, as well as the quality of the data and the metadata" (SeaDataNet, 2010).

Data Transfer

To ensure secure transfer of data, all files will be encrypted using the Advanced Encryption Standard (AES) algorithm. File transfer will be performed via secure protocols such as SSH or SFTP to ensure that no data is intercepted by unauthorized parties.

Data Recovery

In the event of data loss or corruption, provisions will be in place for data recovery. Backups of the server will be created regularly, and the data will be stored off-site in a secure location. In addition, the data recovery process will be tested periodically to ensure its effectiveness.

9. ETHICS

Under the NAPSEA project several interviews and workshops will be conducted. Several ethical key aspects were identified in this respect. In the process of conducting interviews and workshops, collecting data, NAPSEA will ad-here to the ethical principles: the principle of proportionality, the right to privacy, the right to protection of personal data, the right to physical and mental integrity of a person, the right to non-discrimination and the need to ensure high levels of human health protection.

The NAPSEA team is also committed to upholding high standards of academic integrity and will ensure that all sources referenced in the project are properly cited. NAPSEA will adhere to all relevant referencing guidelines and best practices. Accurate referencing is essential to the credibility and legitimacy of the project's findings, and all necessary steps will be taken to avoid any issues of plagiarism or academic misconduct.

10. OTHER ISSUES

NAPSEA won't make use of other national/funder/sectorial/departmental procedures for data management.

11. REFERENCES

Büttner O, Jawitz JW, Birk S, Borchardt D. Why wastewater treatment fails to protect stream ecosystems in Europe. Water Research 2022; 217: 118382.

Cruz-Alonso V, Pucher C, Ratcliffe S, Ruiz-Benito P, Astigarraga J, Neumann M, et al. The easyclimate R package: Easy access to high-resolution daily climate data for Europe. Environmental Modelling & Software 2023; 161: 105627.

Ebeling P, Kumar R, Lutz SR, Nguyen T, Sarrazin F, Weber M, et al. QUADICA: water QUAlity, DIscharge and Catchment Attributes for large-sample studies in Germany. Earth Syst. Sci. Data 2022; 14: 3715-3741.

EEA (2019). https://www.eea.europa.eu/data-and-maps/data/waterbase-uwwtd-urban-waste-water-treatment-directive-7

FGG Elbe 2005 Zusammenfassender Bericht der Flussgebietsgemeinschaft Elbe uber die Analysen nach Artikel 5 der Richtlinie 2000/60/EG (A-Bericht), Tech. rep., Flussgebietsgemeinschaft Elbe, Magdeburg.

ftp://palantir.boku.ac.at/Public/ClimateData/v3/AllDataRasters

https://www.ikse mkol.org/fileadmin/media/user_upload/E/06_Publikationen/08_IKSE_Flyer/2016_ICPER-Flyer_The_Elbe_River_Basin.pdf

https://www.tereno.net/

https://www.ufz.de/moses/

Kumar, R., L. Samaniego, and S. Attinger (2013): Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, Water Resour. Res., 49.

Nguyen TV, Sarrazin FJ, Ebeling P, Musolff A, Fleckenstein JH, Kumar R. Toward Understanding of Long-Term Nitrogen Transport and Retention Dynamics Across German Catchments. Geophysical Research Letters 2022; 49: e2022GL100278.

Samaniego L., R. Kumar, S. Attinger (2010): Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resour. Res., 46, W05523.

SeaDataNet (2010). Quality control procedures.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18

Yang S, Bertuzzo E, Büttner O, Borchardt D, Rao PSC. Emergent spatial patterns of competing benthic and pelagic algae in a river network: A parsimonious basin-scale modeling analysis. Water Research 2021; 193: 116887.