Data enhanced simulation of wakes for all wind turbines in the North Sea nudged by SCADA data deployed on cloud.

A Data Management Plan created using DMPonline.be

Creators: Olivier Ndindayino, Johan Meyers

Affiliation: KU Leuven (KUL)

Funder: Vlaams Agentschap Innoveren & Ondernemen (VLAIO)

Template: VLAIO cSBO DMP (Flemish Standard DMP)

Principal Investigator: Johan Meyers

Data Manager: Olivier Ndindayino, Johan Meyers

Project Administrator: Johan Meyers

ID: 202042

Start date: 01-04-2023

End date: 01-03-2026

Project abstract:

In large wind farms, the wake effect and blockage effect are underestimated problems. The wind is slowed down during inflow, resulting in lower electricity production than expected.

Wind farm design, scenario analysis by governments, grid stability studies, hydrogen wind and energy island design need accurate wake models or digital twin models that are applicable to entire concession zones and make it possible to process monitoring data. The existing models use hyperparameters to do so.

Cloud4Wake will develop methods to define the hyperparameters on the basis of large data sets and thus optimise the accuracy of the models. These data for the North Sea are currently collected from various sources: LIDAR at various locations, meteorological data and data from offshore wind farms.

Intended project results:

1. A new method to estimate wake effects and blockage in a model for a zone of various offshore wind farms. This way, losses due to wake effects can be better estimated by the industry;
2. A cloud-based framework to calibrate these models on large (field) data sets. It is important that farms collecting field data will use them to optimise the design of their wind farms.

Last modified: 09-10-2023

Data enhanced simulation of wakes for all wind turbines in the North Sea nudged by SCADA data deployed on cloud.
VLAIO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, ... • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Simulation Data	input (setup files, initial flow fields, precursor data...), output (computed velocity fields and profiles, flow statistics...) and metadata produced by SP-Wind simulations	Generate new data	Digital	Simulation data	.setup (setup files), .dat (flow fields), .txt (logs), .exe (executables)	• 3D flow field (input and output, always generated, e.g. BL_field.dat): ~1-10GB (depending on the grid size) • 2D planes (output, e.g. time series of velocity planes): ~1-100GB (depending on the grid size) • Precursor simulations: ~GB-TB (depending on the grid size) • Others: ~MB	
SCADA data	Experimental data from windfarms	Reuse existing data	Digital	Experimental	NA	NA	
Post-processing scripts	Python scripts to perform data analysis on the simulation data	Generate new data and Reuse existing data	Digital	software	.py (Python)	~100MB	
Publications	data related to publication papers	Generate new data	Digital	text and figures	.pdf (final pdf), .tex (LaTex files), .png/.jpg (figures), .doc (word document)	~1GB	
Presentations	meeting and conference presentations	Generate new data	Digital	presentations and figures	.ppt (PowerPoint), .pdf, .jpg/.png (figures), .doc (word document), .tex (LaTex files)	~1GB	
Literature	books and papers	Reuse existing data	Digital and physical	textual data, papers, books	.pdf	~1GB	~100 books (in the department)

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- Simulation Data :

Input data (the case setup) is case specific and generated by the user,
templates are included in the SP-Wind code repository. Output data and metadata are generated by the SP-Wind simulation code.

- SCADA data :

Originating from wind farm owners and operators

- Post-processing scripts :

Py4sp python code developed within the TSFO group accessible on the KU Leuven gitlab

- Publications :

Figures obtained from analysis of simulation data through post-processing scripts

- Presentations :

Own presentations and figures obtained from analysis of simulation data through post-processing scripts

- Literature :

Online academic sites: Limo (https://limo.libis.be/index.html) google scholar (https://scholar.google.com/), library,
(physical) books from the TME department

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Yes, Intellectual property rights on SCADA data and other data coming from wind farm owners and operators.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

- Simulation Data :

Each simulation is assigned a distinctive code, which are linked to an individual folder identified by the same code. This folder contains the necessary files for the simulation, including case setup files, source code or executable (for replication), and the output data.
In order to monitor and document the simulations, an Excel spreadsheet "Simlist.xlsx" is used to record all simulation details. This spreadsheet serves as a repository for metadata related to each simulation, organized in accordance with the Dublin Core general metadata standard (outlined below). Additionally, within each simulation folder, there is a README file that provides metadata specific to the simulation. This README is automatically generated, through python scripts, from the case setup files and is further manually updated with metadata according to the Dublin Core metadata standard. The Simlist.xlsx logbook is generated automatically, through python

scripts, based on the case setup files and README.
This simulation logging system was put in place by a PhD student within the department and is accessible on the KU Leuven GitLab server.

- Post-processing scripts :

The Python scripts feature a header that provides an overview of the script's purpose, accompanied by explanatory comments interspersed within the code.

- Publications

Metadata is included in the publications.

- Presentations

Metadata is included in the presentations.

- Literature

Papers are organized by topic, using the Mendeley reference manager.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Metadata for the performed simulations is organized according to the Dublin Core
metadata standard, adapted for our purposes.
Applied to our simulations, the README in the simulation folder (containing the modified
Dublin Core metadata) has the following entries/sections:

- **General information:** cf. Dublin Core (to be filled in manually, this is the responsibility of the person running the simulations)
    - Project: project name
      Description: description
      Purpose: the purpose of the simulation
      Initialized from: relation with previous simulations
      Relevant output: list of (relevant) simulation output files and their file format
      Findings: (brief) discussion of the most important results
      Author: person who runs the simulation
      Date submitted: date the simulation was submitted
      Horizon: recommendations on how long the data should be stored
- **Case setup**: metadata on the case setup, e.g. grid resolution, boundary conditions... (automated)
- **Git info**: git branch and commit hash of the code used in the simulation (automated,
  important for reproduction)
- **VSC information**: start time and end time of the simulation, summary of used resources
  (from the Vlaams Supercomputer Center)

All this metadata is also included in the logbook Simlist.xlsx that contains an overview of all
simulations. The logbook also stores the location of the simulation folder.

3. Data storage & back-up during the research project

Where will the data be stored?

The Vlaams Supercomputer Centrum (VSC) provides a high-performance computing (HPC) supercomputing platform for the simulations. Additionally, they offer complimentary storage infrastructure on their clusters:
$VSC_DATA: hourly backup, 75 GB, permanent storage
$VSC_SCRATCH: 5TB, data deleted 28 days from last access
$VSC_SCRATCH_NODE: 200GB/node available only at runtime (while running a simulation)
Staging (/staging/leuven): permanent storage, the research group uses approx 120TB
shared among all researchers within the group, no backup.

- Simulation Data

Simulations run on the HPC cluster, in their scratch folder on
$VSC_SCRATCH. Runtime data generated during the simulations (e.g. intermediary flow
fields, ~GB-TB) are stored in the scratch folder ($VSC_SCRATCH), if <~4TB. Scratch storage is limited, in volume (<5TB) and time (28 days after last access). Therefore, when simulations are finished:
1. The folder is cleaned: unimportant data (data that will never be reused or data that can be easily reproduced) is removed.
2. The cleaned folder is stored on the TFSO staging folder (/staging/leuven) on the HPC cluster. This is the case for important simulations (important results, validations...).
Or it is stored on $VSC_DATA. This is the case if the simulations do not contain important results,
are limited in volume and/or will be quickly reused.
The metadata of the simulations is stored within the README file located in the simulation folder, as well as in the comprehensive Excel spreadsheet, Simlist.xlsx. This spreadsheet is stored in $VSC_DATA, with an additional copy stored on the PC. Significant results and graphical representations are summarized in Powerpoint or LaTex files in a dedicated documentation folder on the PC.

- Postprocessing scripts

Postprocessing scripts are stored on the HPC cluster, in $VSC_DATA. In that way, postprocessing can happen on the HPC cluster, such that there is no need to copy (potentially very large) data sets to the PC.

- Publications

Data concerning publications are stored in a personal KU Leuven OneDrive folder.

- Presentations

Presentations are stored in a personal KU Leuven OneDrive folder.

- Literature

(Digital) papers and books are stored in a personal KU Leuven OneDrive folder, and they are managed via Mendeley Reference Manager.

How will the data be backed up?

Several storage options outlined above provide automatic backup as a service:
The VSC provides (hourly) backups for data in $VSC_DATA. Data stored on remote git servers is backed up by Git itself. KU Leuven Onedrive folder includes version history.
Note that staging and scratch on the HPC cluster are not back-upped. Since scratch storage is temporary, important data there is always immediately copied elsewhere (as outlined above).

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- No

Staging storage on the HPC cluster (/staging/leuven) may not suffice, since it is shared among all researchers of the TFSO group. In the case of insufficient staging:
First, MANGO will be used (this is a KU Leuven storage system for active research data).
Secondly, the staging and archiving folders will be cleaned (individually by all TFSO researchers), removing unimportant/redundant files. The simulation logbook Simlist.xlsx provides a useful utility to organize file removal in a structured manner. The logbook contains metadata describing how long each simulation data should be kept (so by scrolling through the logbook, researchers can find simulations past their storage horizon which can therefore be safely removed), as well as the amount of resources the simulation consumed (so by scrolling through the logbook, researchers can find simulation data that can be removed since the simulation is easily reproducible).

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Regarding the HPC infrastructure at the HPC:
All data in $VSC_DATA and $VSC_SCRATCH are only accessible by the the user of the VSC-account. It is password protected via a personal private key (public/private key pairs were generated in order to get access to the VSC)
Staging data is accessible for all TFSO group members, via their own VSC account
(that is also password protected, public/private key infrastructure)
Data stored on personal laptop and personal OneDrive is only accessible by the user (password protected).

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Basic storage and backup infrastructure (as well as scratch extension) at VSC is costless
(75GB $VSC_DATA, 5TB $VSC_SCRATCH, 200GB/node $VSC_SCRATCH_NODE).
Staging cost at HPC (/staging/leuven): €20/TB/year * 120TB = €2400/year (for the whole TFSO research group). Cost is carried by the TFSO group.
GitLab, KU Leuven OneDrive, storage on personal laptop are costless.
There are possible costs related to larger data sets stored on MANGO and KU Leuven RDR (~30 euro/Tb/year).

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Only data related to publications or important benchmark/validation data will be preserved. The code generated during the PhD is also preserved in the corresponding git repositories. This data will be kept for a period of minimally 10 years after the end of the project.

- Publications

All data related to publications will be preserved.

- Presentations

Relevant presentation are preserved: conference presentations, presentations serving as documentation for major code developments.
Other presentations - e.g. for TFSO group meetings, weekly meetings with promotor... - are not preserved.

- Literature

Not preserved (except for the physical material that is kept at the department).

Where will these data be archived (stored and curated for the long-term)?

- Simulation Data

After the project, the simulation data in the staging folder is examined:
First, research data that can easily be reproduced will be deleted. Clear
instructions on how to reproduce the simulations (e.g.: which code to use) based on the README in the simulation folder, will be provided.
A selection of the data that will be kept in the staging folder is made. This data, easily
accessible on the HPC cluster, may be used in the future by other researchers in the
TFSO group. This includes benchmark simulations and important research data (that
can serve as a mean to validate future research, serve as comparison...), but also
simulation data that took a very long time to compute such as extensive precursor
simulations (as this will facilitate future research).

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Staging cost at HPC (/staging/leuven): €20/TB/year * 120TB = €2400/year (for the whole
TFSO research group). The cost is carried by the TFSO group.
GitLab repositories are costless.
There are possible costs related to larger data sets stored on MANGO and KU Leuven RDR (~30 euro/Tb/year).

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- No (closed access)
- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, ...)

If access is restricted, please specify who will be able to access the data and under what conditions.

Only KU Leuven personel on a need to know basis and compliant with relevant NDAs.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights

SCADA data and other data coming from wind farm owners and operators.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

In an Open Access repository
In a restricted access repository
Upon request by mail

- Simulation Data

Relevant simulation data is accessible for TFSO researchers via the staging folder of the HPC cluster and MANGO data repository.

- Postprocessing scripts

The postprocessing scripts will be integrated in a general postprocessing framework for all SP-Wind users, stored on a git repository. The scripts are accessible from there (for future TFSOresearchers).

- Publications

All data related to publications (as explicited in the previous sections) will beaccessible. The manuscripts of the publications is also accessible via Lirias (open access). For as far as this is possible for data size all scripts and data will be shared through KU Leuven RDR.

When will the data be made available?

At the time of publication.
All data that will be shared will be available immediately (as soon as they are stored on the archive folder of the HPC cluster, the shared folder, git...).

Which data usage licenses are you going to provide? If none, please explain why.

GNU license with no commercial use or freer

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

When published on KU Leuven RDR.

What are the expected costs for data sharing? How will these costs be covered?

There are possible costs related to larger data sets stored on MANGO and KU Leuven RDR (~30 euro/Tb/year).

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Olivier Ndindayino and Johan Meyers

Who will manage data storage and backup during the research project?

Olivier Ndindayino and Johan Meyers

Who will manage data preservation and sharing?

Johan Meyers

Who will update and implement this DMP?

Olivier Ndindayino and Johan Meyers

Data enhanced simulation of wakes for all wind turbines in the North Sea nudged by SCADA data deployed on cloud.
GDPR

GDPR

Have you registered personal data processing activities for this project?

- No

Data enhanced simulation of wakes for all wind turbines in the North Sea nudged by SCADA data deployed on cloud.
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable