# Computational methods for infinite-dimensional Bayesian inversion of physics-based models in engineering applications

*A Data Management Plan created using DMPonline.be*

**Creator:** Giovanni Samaey https://orcid.org/0000-0001-8433-4523

**Affiliation:** KU Leuven (KUL)

**Template:** KU Leuven BOF-IOF

**Principal Investigator:** First Name Surname https://orcid.org/0000-0002-9273-3038, Giovanni Samaey https://orcid.org/0000-0001-8433-4523, n.n. n.n., Johan Meyers

**Grant number / URL:** https://research.kuleuven.be/portal/en/project/3E230222

**ID:** 206137

**Start date:** 01-10-2023

**End date:** 30-09-2028

**Project abstract:**

In this proposal, we will develop particle methods for sampling and optimization in computationally challenging Bayesian inverse problems. We will consider problems that are modeled by a (possibly high-dimensional) PDE that describes evolution of a system in space, time and potentially contains additional degrees of freedom. In this model, we need to estimate unknown parameters that are potentially infinite-dimensional (e.g., a function of space), based on measurement high-resolution measurement data. The main objectives of this proposal are three-fold. We will increase reliability of the computed Bayesian posterior distributions quantifying and reducing bias due to model error, and by quantifying uncertainty due to noisy measurement data. We will moreover increase efficiency of computation by exploiting their multilevel/multiscale structure in the computational framework. Finally, we will establish a link between Bayesian inverse problems and deep learning, opening an additional path towards explainable AI.

**Last modified:** 02-04-2024

# Computational methods for infinite-dimensional Bayesian inversion of physics-based models in engineering applications

### 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.**

This project will generate new data and re-use existing data.
Three types of data:

- own algorithms and in-house software that is being developed during the project
- simulation results, obtained using either commercial software, open source software, or in-house codes.
- processed simulation results (figures, tables, etc.)

The specific data format depends on the code being used. This will be documented with each computational experiment, see section 4 (documentation and metadata).
The expected volume of the data is hard to measure at this point.  Since the goal of the project is the development of new simulation tools, we expect large amounts of data, generated by the algorithms we develop, with different life spans:

- simulation results that are meant to test and optimize new methods and codes (requiring only short-term storage)
- simulation results that will be used to support claims in scientific papers (requiring long-term storage).

**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:**

Re-used data will consist of simulation results generated by our own software in previous or related projects.

**Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.**

- No

**Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).**

- 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.**

- Yes

Although tech transfer and valorisation are not direct goals of this project, the developed software and algorithms might yield opportunities for valorisation. Whenever such an opportunity presents itself, it will be discussed in the project’s steering committee. The conclusions of that discussion will be appended to this data management plan.

**Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or 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.**

- No

**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

### 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).**

We require that each data file is accompanied by instructions on how to store, open and read it:

- Data and scripts should be stored together in the same folder.
- All data fields should have meaningful names.
- All data files should be accompanied by a README file that describes the goal of the experiment, the data format and the meaning of all stored quantities.
- Each README file should contain the names of the data files and script files, as well as the (version of) the software that generated it.
- There should be a relation between the content of the experiment and the name of the script.
- There should be a relation between the script name and the name of the derived data files and figures. For figures, ensure that the plotted quantities are in the file name.
- If the script takes input parameters, these should be used in the name of the data files.

We are currently migrating to a MANGO setup (KU Leuven's in-house data management system), which will be integrated with the workflow of the researchers to automate the above procedures and produce "computational experiment" records. These records can then easily be shared via KU Leuven's RDR system upon finalizing publications.

**Will a metadata standard be used to make it easier to find and reuse the data?**
**If so, please specify which metadata standard will be used.**

**If not, please specify which metadata will be created to make the data easier to find and reuse.**

- Yes

Although no formal metadata standards are applicable to this type of research, the standardized way in which all steps in the research will be documented (see previous question) will ensure findability and reusability of the data.

### Data Storage & Back-up during the Research Project

**Where will the data be stored?**

- ManGO
- OneDrive (KU Leuven)
- Shared network drive (J-drive)
- Personal network drive (I-drive)
- Large Volume Storage

- Other (specify below)

Other:

- gitlab.kuleuven.be
- departmental storage systems in the involved departments
- VSC archive storage

This project is a collaboration between research groups of different departments. Each of these groups has a policy on how data should be stored.

For software:

- Software is stored in a version control repository.  We choose git.
- Each software package has its own repository, with dedicated rules on access, commit procedures, branching, etc. Software is at least accessible to all researchers that contribute to a specific manuscript.
- Each involved department has internal procedures for the git server.

For manuscripts:

- Manuscripts are also stored in a version control repository, based on git.
- Each manuscript has its own repository, at least accessible by all co-authors of the manuscript.
- The choice of git server will be made on a case-by-case basis.

For simulation data:

- Data is stored either on the personal computer of the researcher or on the supercomputer on which computations were done. (E.g. using archive nodes on the VSC.)
- From there, data is synced to a place that is accessible to the PIs of this project and collaborators.  This can be done using the KU Leuven OneDrive system, syncing with a departmental file server, or similar.

**How will the data be backed up?**

- Standard back-up provided by KU Leuven ICTS for my storage solution

Once data is synced to university machines (either via OneDrive or via a departmental file server), the university systems for backup are used. OneDrive ensures backup. Departments with departmental file servers (such as Computer Science) guarantee that their file servers are backup up correctly.
Once each researcher syncs with the centrally controlled system, backup is automatic.

**Is there currently sufficient storage & backup capacity during the project?**

**If no or insufficient storage or backup capacities are available, explain how this will be taken care of.**

- Yes

We will systematically buy storage space as required to fulfil the backup needs of the project.

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

Each researcher is responsible for the security of his/her personal computer, by keeping operating system and software up to date.
For data that is stored on version control systems or on centrally maintained systems, we rely on ICTS for security.
Computational experiments that are pushed to KU Leuven ManGo become read-only to avoid inadverted modifications or inconsistencies between algorithm and simulation outcomes.

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

The costs for data storage result from university policy and are documented on the pages of ICTS. We have explicitly foreseen a working budget with this C1 proposal to cover these costs.

### Data Preservation after the end of the Research Project

**Which data will be retained for 10 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...).**

- Certain data cannot be kept for 10 years (explain below)

- All software and scripts will be retained in the version control system.
- Not all simulation data files can be retained, due to physical storage constraints. However, all data files that are required to generate figures in published papers will be retained for at least 10 years after the end of the project.

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

- KU Leuven RDR
- Large Volume Storage (longterm for large volumes)
- Shared network drive (J-drive)

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

The costs for data storage result from university policy and are documented on the pages of ICTS. We have explicitly foreseen a working budget with this C1 proposal to cover these costs.

### Data Sharing and Reuse

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

- Yes, as open data

Whenever appropriate, all data and code required to reproduce figures and tables in papers will be made openly accessible.
In cases where this would not be possible, data will be made available upon request to the scientific community.

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

Members of the scientific community, upon request. Specific licence agreements will be discussed when such requests occur.

**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 per dataset or data type where appropriate.**

- No

**Where will the data be made available?**

**If already known, please provide a repository per dataset or data type.**

- KU Leuven RDR (Research Data Repository)

**When will the data be made available?**

- Upon publication of research results

**Which data usage licenses are you going to provide?**

**If none, please explain why.**

- MIT licence (code)
- GNU GPL-3.0 (code)
- CC-BY 4.0 (data)

The specific license will be decided case by case, as this requires an agreement between all authors.

**Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.**

- Yes, a PID will be added upon deposit in a data repository

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

The costs for data storage result from university policy and are documented on the pages of ICTS. We have explicitly foreseen a working budget with this C1 proposal to cover these costs.

## Responsibilities

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

A data management guide has been written to inform all researchers on how to manage the data they generate. Each PI monitors the implementation of these guidelines for his/her own researchers.

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

A data management guide has been written to inform all researchers on how to manage the data they generate. Each PI monitors the implementation of these guidelines for his/her own researchers.

### Who will manage data preservation and sharing?

The PIs of this C1 project have regular steering committee meetings, in which requests to share data with external scientists will be discussed.

### Who will update and implement this DMP?

The PIs of this C1 project have regular steering committee meetings, in which also this data management plan will be monitored and updated.