DMP - Surrogate Modeling of the Plasma Environment of magnetized and un-magnetized rocky bodies

Project Name My plan (FWO DMP) - DMP - Surrogate Modeling of the Plasma Environment of magnetized and un-magnetized rocky bodies

Project Identifier 3E220064

Grant Title 1SF8522N

Principal Investigator / Researcher Hanne Baeke

Project Data Contact hanne.baeke@kuleuven.be

Description My Ph.D. thesis aims to improve the knowledge of space weather conditions around magnetized (Mercury) and un-magnetized (Moon) rocky bodies, by studying the multi-scale plasma dynamics of Mercury and the Moon, two bodies in close interaction with the solar wind. Therefore, kinetic particle-in-cell simulations will be performed for a variety of solar wind conditions. These will be gathered in a dictionary, that will be used to construct a machine learning surrogate model. Such model is capable of replacing the simulations and allows for an almost instantaneous computation of the present 3D plasma environment under given (observed) solar wind conditions. The surrogate model can be included as space weather prediction tool in future Lunar missions.

Institution KU Leuven

1. General Information

Name applicant

Hanne Baeke

FWO Project Number & Title

1SF8522N

SUrrogate modelling of the Plasma Environment of magnetized and un-magnetized Rocky Bodies (SUPERB)

Affiliation

- KU Leuven

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Generate new data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

This project will focus on the development of a machine learning surrogate model for kinetic particle-in-cell (PIC) simulations of the plasma environment around Mercury. For this, the kinetic PIC code ECsim is used to generate the training data for the ML algorithm.

Generated data:

Type of data	Format	Volume	How created
2D and 3D simulations field data	matrices for up to 15 properties in HDF5 format	5 TB	Simulations are performed using ECsim, a kinetic scale plasma simulation code.
2D and 3D simulations particle data	1D arrays containing position and velocity for up to 1 billion particles per simulation in HDF5 format	35 TB	Simulations are performed using ECsim, a kinetic scale plasma simulation code.
Code of machine learning surrogate model	python scripts		A large part of the project will go to the development of the algorithm for the surrogate model. This will be written in Python, with PyTorch as ML package.

Depending on how the project evolves, the algorithm might be tested on real spacecraft data, from the Bepicolombo mission. This data is yet to be collected, so the format and volume is still unsure.

Existing data:

Type of data	Format	Volume	How accessed
Spacecraft data Bepicolombo		1 TB	through ESA website

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

NA

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, add the reference to the formal approval by the relevant ethical review committee(s)

- No

NA

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- No

NA

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

NA

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

1. The simulation code ECsim contains a folder 'Documentation'. Inside this folder there is for example an output.txt file that tells the user how to read the output files. There is also a units.txt file, which tells the user which units are used for the different physical parameters.

The simulation data will be organized in folders with meaningful names, one per simulation run. Each folder will contain a README.txt, explaining how to rerun the specific case and what the related input and output files are. All the output files will be organized in a 'data' folder within the correct folder. The simulation code also generates a SimulationData.txt file in 'data', which summerizes the settings in the input file of the specific simulation. Visualization output will be organized inside a 'png' folder with a README.txt containg information on what is visualized and how.

Centrally, there will also be a README.txt file containing the physical parameters that are outputted by the simulation code ECsim.

2. The python scripts for the surrogate model will be published on github together with a README.txt file, explaining what is inside each script and used data file and how to run the algorithm (training, evaluation, generating results).

Examples of parameters that will be inside the data filles for training: position of particles, magnetic field, timestep, etc.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

Part of the metadata is automatically generated by the simulations (SimulationData.txt). This file describes already all the information needed to reproduce the simulation. In addition I will add my own README.txt stating the purpose for which this run was done, whether troubles were encountered, how this will be processed and which version/branch of the code was used to run the simulation.

5. Data storage and backup during the FWO project

Where will the data be stored?

1. Simulations are currently run on SuperMUC-NG (LRZ), which is also where the output data is temporarily stored.

There is a multi-level storage solution implemented to manage the team's data, including a local NAS, shared disk space at VSC, and a freely accessible data repository for publications.

The shared disk space at VSC will be used for data that needs to be shared efficiently within the team. The rest of the data will be stored at the local NAS.

2. The code for the machine learning surrogate model will be stored in a GitLab repository. This includes the source code, input data, intermediate results and evolution of the scripts.

How is backup of the data provided?

1. The data saved on my work laptop is stored inside my KU Leuven OneDrive repository and is automatically backup up to the OneDrive two times a day.

The simulation data will be stored and backed-up at the local NAS of the research group. The most important data will also be stored at the VSC staging repository of the research group.

2. All the data uploaded to the GitLab repository will be automatically backup under versioning control.

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.

- Yes

The storage and backup capacity is sufficient. The largest data are the particle files. These are mainly needed for restarts, so only the file of the latest timeframe is kept. The fields data is kept for more regular time frames. The storage in the local NAS has a capacity of 100 TB, which can be extended if needed.

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

The amount of data that will be produced during the PhD is estimated to be around 40 TB. The largest part of the data will be stored at the NAS, which was a one-time expense at purchase time, so there are no specific costs connected to that storage.

The VSC staging repository will be used for data that needs to be accessible to other members of the group. The VSC staging repository costs € 20 per TB, per year, and needs to be payed only for the amount of storage that is used. I estimate the amount of data to be stored at the VSC at a maximum of 5 TB. This would require a maximum of € 100 per year.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

I will not be working with any sensitive data, so this issue is not really a concern. After publishing, the data and code used will be made publicly available through a public GitLab repository.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

All data leading to publications will be stored to guarentee reproducibility for at least 5 years after the end of the project.

Due to the amound of data, maintaining all particle data produced by each simulation would be impossible. We do strive to maintain all essential data generated and therefore, we will limit the storage of particle data to the principal simulations of this project and to those that lead to scientific publications.

Moreover, only one particle file, used for data analysis, will be preserved per simulation, to avoid I/O bottlenecks in the data analysis.

Finally, initial and boundary conditions will be maintained for reproducibility.

One of the main aims is the creation of a dictioary of very detailed and high resolution simulations. This dictionary (training data for machine learning model) will be preserved for more than 5 years.

Where will the data be archived (= stored for the longer term)?

A small part of the data will be archived on the GitLab repository (python scripts, output examples, test data set).

The simulation results will be archived on the NAS of the research group. Other researchers can get access to this data upon request by email. In the future an iRODS will be implemented on the NAS, in order to share the data with outside researchers more conveniently.

The RDR of KU Leuven will be used to store the most essential data connected to publications (up to 50 GB per year). This data will be accessible by interested researchers.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

The storage of data in the RDR of KU Leuven is free up to 50 GB per year. The costs for the NAS are a shared one-time expense of the research group as a whole, so there are no specific costs involved with that storage unit.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No

NA

Which data will be made available after the end of the project?

1. The full training data set used for the machine learning surrogate model will be made available. Part of it will be freely available for download. If one would want to access the complete data set, this will be possible upon request by email.

2. The python scripts for the machine learning algorithm will be publicly available in a GitLab repository.

Where/how will the data be made available for reuse?

- In an Open Access repository
- Upon request by mail

The scripts, example output and some test data will be available through the public access GitLab repository. Part of the data set will be available through RDR from KU Leuven (up to 50 GB). Upon request by email, access to the full data set will be granted as well.

When will the data be made available?

- Immediately after the end of the project
- Upon publication of the research results

The data used for publishing will be made available together with the research publication.

Who will be able to access the data and under what conditions?

The data will be available to anyone for any purpose, provided that they give appropriate credit to the creators.

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

There are no expected costs for data sharing. The RDR of KU Leuven is free up to 50 GB per year. The use of GitLab is free as well. The rest of the data will be stored at NAS, no expenses are connected to this as well.

8. Responsibilities

Who will be responsible for data documentation & metadata?

Hanne Baeke

Who will be responsible for data storage & back up during the project?

Hanne Baeke

Who will be responsible for ensuring data preservation and reuse ?

Hanne Baeke

Who bears the end responsibility for updating & implementing this DMP?

The PI and her supervisor bear the end responsibility of updating & implementing this DMP.