DMP title

Project Name (Machine) Learning from our Mistakes: Confronting Land Surface Models with Artificial Intelligence - DMP title

Project Identifier KU Leuven BOF C14/21/057

Grant Title C14/21/057

Principal Investigator / Researcher Gabrielle De Lannoy

Project Data Contact Gabrielle.DeLannoy@kuleuven.be

Description Theory-based land surface models (LSMs) represent our current best knowledge about biogeophysical land surface processes and are being used for weather forecasting, climate modeling, and a host of other environmental applications. Data-driven machine learning (ML) techniques have become ubiquitous in almost all fields of research as well as in everyday life, yet its potential remains underexploited in many fields of the Earth sciences. This project aims (i) to harness ML to better understand the connections between various land surface variables (in particular soil moisture, vegetation, and snow) and their meteorological drivers (such as precipitation, temperature, solar radiation, and wind), and (ii) to confront ML with modern LSMs and satellite-based land data assimilation to further our knowledge about biogeophysical processes, thus fostering the development of advanced physical theories and a ''digital twin earth''.

1. General Information

Name of the project lead (PI)

Gabrielle De Lannoy

Internal Funds Project number & title

C14/21/057 (Machine) Learning from our Mistakes: Confronting Land Surface Models with Artificial Intelligence

2. Data description

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

- Generate new data
- Reuse existing data

2.2. What data will you collect, generate or reuse? Describe the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a numbered list or table and per objective of the project.

Existing data*:

Type of data	Format	Volume	Origin
Sentinel-1 satellite data	Netcdf	10 TB	ASF, ESA
Sentinel-1 based snow depth	Netcdf	100 GB	VSC HPC
SMOS based vegetation optical depth	Binary	2 TB	ESA
SMAP based vegetation optical depth	Netcdf	2 TB	NASA
Satellite-based LAI data	Netcdf	2 TB	NASA
MERRA-2 forcing data	Netcdf	20 TB	NASA
Ancillary data for land surface modeling	Binary files	10 GB	NASA

* Abbreviations: ASF - Alaska Satellite Facility; VSC HPC - Vlaams Supercomputer Centrum High Performance Computing; MERRA2 - Modern-Era Retrospective analysis for Research and Applications, Version 2

Generated data:

Type of data	Format	Volume	Origin
Land surface model output	Netcdf	5 TB	VSC HPC
Machine learning output	Netcdf	500 GB	VSC HPC

3. Ethical and legal issues

3.1. Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

No

3.2. 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

3.3. Does your research 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

3.4. 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 regarding reuse and sharing are in place?

No

4. Documentation and metadata

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

All generated model output is produced in a standardized and documented way, either with metadata available inside the files, or via user documentation. Most model infrastructure will also be documented on GitHub.

4.2. Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.

Partly. For details on metadata standards produced by the modeling framework, please see https://github.com/GEOS-ESM/GEOSldas

5. Data storage and backup during the project

5.1. Where will the data be stored?

All data will be stored on the VSC HPC large volume storage.

5.2. How will the data be backed up?

Automatic backup on the VSC HPC.

5.3. 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, i.e. paid storage on the HPC (/staging and /archive) and Storage4Climate.

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

~5000 Euro, costs will be covered by the project.

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

Access to the HPC is password and key protected.

6. Data preservation after the end of the project

6.1. Which data will be retained for the expected 10 year period after the end of the project? If only a selection of the data can/will be preserved, clearly state why this is the case (legal or contractual restrictions, physical preservation issues, ...).

Final generated data used in publications will be preserved.
Re-used data will not necessarily be preserved, unless they remain used or are no longer supplied by the original data providers.

6.2. Where will these data be archived (= stored for the long term)?

VSC HPC (/archive).

6.3. What are the expected costs for data preservation during these 10 years? How will the costs be covered?

The final output of this project will be limited compared to all storage needed for the project, and is estimated at ±1000 euro for 5 years. The cost will be covered by future projects, and the promotor's basic working fund.

7. Data sharing and re-use

7.1. 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 or because of IP potential)?

No

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

Sentinel-1 based snow depth retrievals

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

- In an Open Access repository

7.4. When will the data be made available?

- Upon publication of the research results

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

The data will be accessible to everyone.

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

None

8. Responsibilities

8.1. Who will be responsible for the data documentation & metadata?

Postdoctoral researcher to be hired.
Gabrielle De Lannoy, head of research group.

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

Gabrielle De Lannoy, head of research group.

8.3. Who will be responsible for ensuring data preservation and sharing?

Gabrielle De Lannoy, head of research group.

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

The end responsibility for updating and implementing the DMP is with the supervisor (promotor).