Plan Overview

A Data Management Plan created using DMPTuuli

Title: Combining Coupled Modelling and Machine Learning to Constrain Antarctica's Uncertain Future (project number 355572)

Creator: Rupert Gladstone

Principal Investigator: Professor John Moore

Data Manager: Rupert Gladstone

Project Administrator: Rupert Gladstone

Contributor: Rupert Gladstone

Affiliation: University of Lapland

Funder: The Research Council of Finland (former The Academy of Finland)

Template: Academy of Finland data management plan guidelines (2021-2023)

ORCID iD: 0000-0001-8271-5787

Project abstract:

The Antarctic Ice Sheet is losing mass. Large regions of the grounded ice sheet respond to this by accelerating and thinning. This in turn modifies subglacial conditions, hydrology, and glacier sliding, completing a complex three-way set of interactions between ocean, ice, and subglacial water. Increased freshwater flux from Antarctica impacts the surrounding coastal seas and Southern Ocean, changing currents, hydrography and sea level, and hence affecting the global climate system. Sea level rise provides Antarctica's primary societal impact, threatening more than a half a billion people. While the latest technical summary from the Intergovernmental Panel for Climate Change (IPCC) anticipates several tens of centimeters of sea level rise by the end of the 21st century, they also state that "there is deep uncertainty in sea level projections for 2100 and beyond associated with the ice-sheet responses to warming. In a low-likelihood, high-impact storyline and a high CO2 emissions scenario, ice-sheet processes characterized by deep uncertainty could drive GMSL rise up to about 5 m by 2150". Such a high increase in sea level would be globally catastrophic, but how likely is it? We will bring together physical glaciology and oceanography, and state-of-the-art analysis techniques in statistics, high performance computing, and machine learning to address this question.

Our team of researchers in Finland and international collaborators are at the forefront of ice sheet modelling advances, having developed a computer modelling system in which:

• The full stress balance within the ice sheet is represented, capturing the transitions between three key regimes of slow flowing, fast flowing, and floating ice.

• The evolution of water filled cavities and channels beneath the ice sheet are incorporated and interact with the ice sheet, an important feedback on the speed at which the ice sheet slides over its bedrock.

• The interaction between this subglacial water, the surrounding ocean, and ocean-induced melting of the ice are included through coupling to an ocean model using the Framework for Ice Sheet - Ocean Coupling.

Individually, any of the above would make our approach cutting edge. Collectively, no other research team has this capability.

We will use machine learning techniques to create an emulator for our Antarctic model in order to provide probabilistic projections for Antarctica's contribution to sea level over the next two centuries.

ID: 22248

Start date: 01-09-2023

End date: 31-08-2027

Last modified: 15-06-2023

Grant number / URL: Decision number 355572

Copyright information:

Combining Coupled Modelling and Machine Learning to Constrain Antarctica's Uncertain Future (project number 355572)

1. General description of data

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Additionally, give a rough estimate of the size of the data produced/collected.

Input data for the project comprise climate data from observations and computer models. This includes satellite observed ice flow speeds, ice sheet surface conditions from climate models, projected change in ocean temperatures from climate models.
The input data will mostly be in netcdf format.
We estimate a total of between 500GB to 1TB of disk space will be needed to contain input data.
The data produced will be ice sheet model outputs such as projected ice flow speed and ice sheet contribution to sea level over multiple centuries.
The outputs data will be in pvtu format (for analysis in visualisation software Paraview) and netcdf format (for numerical post processing).
We estimate 60TB disk space will be neede to contain raw output data, and 10TB will be needed for long term storage of processed output data.

1.2 How will the consistency and quality of data be controlled?

The consistency and quality of input data will be controlled through assessment of peer reviewed papers describing the data.
The consistency and quality of output data will be controlled through model verification and validation. More specifically, Bayesian updating based on conparison between model outputs and observed data will be used.

2. Ethical and legal compliance

2.1 What legal issues are related to your data management? (For example, GDPR and other legislation affecting data processing.)

We do not believe legal issues apply. The input data are publically available and can be freely used for research purposes. No personal data will be used.

2.2 How will you manage the rights of the data you use, produce and share?

The computer codes used to generate our output data are publically available through online repositories under creative commons licenses such as GPL 2.0.
The raw model outputs will be stored locally on CSC machines and made available to others on request. The more significant processed model outputs will be made publically available through CSC services and or online data repositories.

3. Documentation and metadata

3. How will you document your data in order to make the data findable, accessible, interoperable and re-usable for you and others? What kind of metadata standards, README files or other documentation will you use to help others to understand and use your data?

Details describing the data will be available through supporting materials to these articles and or separate data descriptions made available through Zenodo and assigned a DOI.
Processed output data will be in self-describing netcdf data files, following CF ("Climate and Forecasting") meta data conventions.

4. Storage and backup during the research project

4.1 Where will your data be stored, and how will the data be backed up?

Input data and raw output data will be stored on scratch spaces on CSC machines. These data are all reproducible and will not be backed up.
If we apply significant processing to input data these will will be backed up using CSC's ALLAS service.
Key output data and all processed output data will be stored on CSC's ALLAS service.
A small key subset of these data will be made publically available (ALLAS can provide URLs to data sets).

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

Dr Rupert Gladstone will lead data management. Practical aspects will be administered through CSC, specifically through consultant Thomas Zwinger, who is funded for 7 months of time over the 4 year project.
Unless specifically made publically available, all data on CSC machines is accessible only to project members and international colaborators who have specifically been given access to the project.

5. Opening, publishing and archiving the data after the research project

5.1 What part of the data can be made openly available or published? Where and when will the data, or its metadata, be made available?

Data will be stored using IDA or EUDAT (to be determined).
All output data will be publically available in principle, though we do not anticipate that others will be interested in obtaining the raw output data. Raw output data will be made available on a case by case basis.
Key subsets of the output data related to policy-relevant sea level projections will be made publically available.

5.2 Where will data with long-term value be archived, and for how long?

Key output data will be archived on CSC services for 10 years.

6. Data management responsibilities and resources

6.1 Who (for example role, position, and institution) will be responsible for data management?

Adjunct Professor Dr Rupert Gladstone of University of Lapland will lead data management, with assistance from Adjunct Professor Dr Thomas Zwinger of CSC.

6.2 What resources will be required for your data management procedures to ensure that the data can be opened and preserved according to FAIR principles (Findable, Accessible, Interoperable, Re-usable)?

A combination of public repositories and CSC services are able to provide all necessary resources.
This includes:
git repositories for software.
Zenodo and or supporting material for open access journal articles for dataset descriptions.
CSC's backup and archiving services, including ALLAS and IDA or EUDAT.