Data and Software Sharing Guidance for Authors Submitting to AGU journals

Data and software are the building blocks of the research published in the AGU journals. These digital objects need to be accessible, understandable, and open as possible for reuse to support transparency and replicability. These digital objects include:

1) Data from observations collected in the field;
2) Data from satellites (primarily level 2 or 3);
3) Data from laboratory experiments;
4) Software used for analysis and visualization of the data;
5) Software used to produce model output;
6) All data displayed in the figures of the paper.

Data and Software Availability Statements and Citations must satisfy AGU's Data and Software for Authors requirements before publication. In this document, we offer guidance, templates, and examples to assist authors in meeting these requirements. The final determination of whether a manuscript meets these requirements is made by the journal editors. Author feedback is appreciated to help ensure that the process remains efficient, feasible, and meaningful.

AGU recognizes that not all data or software can be fully open. Data or software that are sensitive or restricted must be protected through appropriate access controls. Data or software should be as open as possible, as closed as necessary. For data concerning Indigenous Peoples, authors should consult the CARE Principles for Indigenous Data Governance.

AGU Data Help Desk
For questions or feedback regarding AGU's Data and Software Sharing Guidance, contact DataHelp@agu.org.

Considerations for publication related to data and software

1. **When to Make Your Data and Software Available**
    At the time your paper is <u>submitted</u>, your data and software must be available to the editors and reviewers. At the time your paper is <u>accepted</u>, your data and software Availability Statements must be clearly stated. A few repositories require the paper to be published before the data being registered and available, such that the persistent identifier resolves. These repositories are known, community-accepted repositories, and acceptable for use. Authors must still provide preliminary access to reviewers at the time of submission. Please ensure that your data and software are available with details of the online access location(s), data product names, variable names, time ranges, spatial locations, or any other search criteria to allow a reader and reviewer to Find and Access the data used and/or generated for the paper (including those represented in figures and tables). In summary, any data and software utilized in the work contained in the manuscript must be documented for free and open availability. Data or software that are sensitive and require restrictions on access (e.g., personal data, medical information, fossil locations, strategic models) must be preserved in a repository with appropriate access controls.

2. **Availability Statement in Open Research Section (required)**
    a. <u>Data Availability Statement (required):</u> A Data Availability Statement is required in the Open Research section of your paper describing where and how your data are available, including an online means to access your data. Check links and files before submitting your paper to the journal so as to ensure the data are accessible for peer review. Many data repositories provide confidential data access for this purpose. <u>[For data that are not publicly available, sensitive, or restricted,](#) [examples, templates, and specific guidance are provided in this document.](#)</u>

    b. <u>Software Availability Statement (required):</u> A Software Availability Statement is required in the Open Research section of your paper for software that is central to your research such as for model simulations, data analysis, data visualization, and model output analysis.
    The Availability Statement should contain a citation, licensing information, access restrictions, and a link to the development platform (e.g. GitHub). Note that "git/GitHub/GitLab" are not acceptable software repositories because they are not archival. <u>[For software that are not publicly available, sensitive, or restricted, examples, templates, and specific guidance, are provided in this document.](#)</u>

3. **Data Citation (required)**

   Include data citations for the primary and processed research data in the References section of your paper. Doing so ensures proper credit is given for the data. Note, English-language (or English translation) for any cited sources is required. The [AGU's Data and Software for Authors](#) guide provides information on how to cite your data.

   a. Primary Data and Processed Data: Some repositories "reserve a DOI" before publishing your data. Use this DOI in your data citation. Once your data are published, the DOI will resolve properly. Some repositories use other persistent identifiers or URLs which are permissible (e.g, GenBank). Some repositories will only provide a DOI close to or at the time of paper acceptance; in this case the DOI will need to be added during the final revision.

   b. Simulated Data / Model Output Data: [See the guidance for Numerical Models.](#)

   c. Large Data (>1TB): Preservation of large data may be possible in some repositories for a fee. Authors should account for this in their research budget and Data Management Plan. Also, check with your university or institution and their repository.

   d. Data Used from Another Source (e.g., data created by others): Cite the specific source to ensure proper credit and allow readers to access the same version of the data. If the data are also associated with a publication, especially a data publication, cite both the paper and the repository.

4. **Software Citation (required)**
   If software is used to analyze or produce the data, including for model output, then include a software citation in the References section of your paper. The [AGU's Data and Software for Authors](#) guide provides information on how to cite your software.

   a. Software you or your team have written: see the section below on preserving software.

   b. Software already preserved: simply cite it.

   c. Software created by others, but not preserved: work with the author on how they want it to be cited.

Guidelines for Research Primarily Based on Numerical Models or Theory

While numerical models or theoretical work may not utilize (input) data, often "output" such as figures or tables are considered data and should be made available in electronic form. Additionally, the software code (e.g. Python, Jupyter Notebooks, R, MATLAB) used to perform any data analysis and to produce the manuscript's figures should be made available in a free and open platform (e.g., Github) and preserved in a repository (e.g., Zenodo). In the case where a manuscript makes no use of models, data, or analysis software (e.g., a purely theoretical paper or a review paper), then make note of this point in the Data and Software Availability Statements.

When the primary data for the research comes from numerical model simulations, follow these guidelines:

1. **Citation of the model software**

   a. BEST OPTION (model in repository): Cite the model using a repository that registers the version used for the paper with a persistent identifier (e.g., Digital Object Identifier) and metadata that describes the model using community standards. For example, Github provides a [connection to Zenodo](#) for this purpose. If a published paper has the complete description, there should be a link in the repository to the published paper. Your citation should accurately capture the authors/creators of the model. In the Ocean modeling community it is common to use numerical models that are open access and well documented (e.g., GFDL-MOM, NEMO, ROMS, ADCIRC, FESOM, SHYFEM, SURF).

   b. GOOD OPTION (model described in paper): Cite the publication where the numerical model is described with information about the version used for this paper.

2. **Description of the numerical model.**

   a. Include a description of the model in the text of the paper that is adequate to support replicability. If a publication describes the model thoroughly, cite that paper.

3. **Information about the configuration/parameters used to run the model.**

   a. This information should be included in the paper text as well as providing any script/workflow used. The script/workflow should be preserved in a repository and cited. Any boundary and/or initial condition datasets used should be described and cited. The goal is to provide sufficient information and resources so that an interested user, with sufficient computer resources, can replicate your simulation.

4. **Data and analysis software that supports the Summary Results, Tables and Figures.**

   a. BEST OPTION: Cite a package in an appropriate repository that includes scripts/workflows, provenance information, and summary files that support the research, figures, and tables, consistent with archives maintained for transparency and traceability by assessments such as the IPCC.

   b. GOOD OPTION: Cite files (e.g., scripts, descriptive detail) in an appropriate repository that support evaluating the research and provide the details behind the tables and figures.

   c. ACCEPTABLE OPTION: Provide the necessary information for transparency and traceability of the analysis using your community standards or guidance.

5. **Model Output Data.**

   e. If model output is instrumental to evaluating the research, particularly with respect to producing manuscript figures or tables, then deposit the necessary model output in a community accepted, trusted repository There are currently limited resources for preserving files of very large size. However, selecting adequate output to produce manuscript figures and tables is generally much more manageable and is sufficient to meet the needs of replicability.

AGU journals strongly prefer the publication of free and open-source software to ensure the replicability of results by readers.

Proprietary or not "freely" available software can be used and cited provided that readers are able to access the software through standard and reasonable means (e.g., a software package associated with an instrument, or an available visualization script). Standard graphics, spreadsheet, or word processing programs do not need to be cited.

Software that can not be made available during peer review may result in the paper not being accepted. The editor must be consulted in this case.

Highlight: When deciding on what model data (e.g., simulation workflow outputs), simulation workflow configuration and code components to include with your paper, refer to the rubric and guidance developed by the [EarthCube Research Coordination Network (RCN)](#) on model data management best practices.

Selecting Your Repository
Selecting a repository and determining data and software management best practices begins when you propose and fund your research project, through your Data (and Software) Management Plan.

For publishing, you need to locate a repository that provides preservation services, such that:

1. The repository registers your data with a persistent identifier that is globally unique such as a Digital Object Identifier (DOI).

2. The data are freely accessible from a landing page that provides information (e.g., metadata) about your data, and preferably version controlled.

Once published in an appropriate ***preservation*** repository, your data cannot be modified. We recommend you be in contact with the repository to understand their preservation practices and how they support the community and journal requirements.

Discipline-Specific Repositories: These are specialized repositories which typically provide support and information on required standards for metadata and more.

Institutional Repository: Many universities are supporting research data and software management and compliance requirements for their researchers, and such services are often provided through the library. Librarians can be an excellent source of research data management support, including repository selection, and can help you comply with funder, publisher, and university requirements.

General Repository: When using a general repository, make sure you provide documentation about your data that is in line with your community standards. Please refer to the Generalist Repository Comparison Chart for additional repositories and guidance. Common repositories are: Zenodo, Dryad, or Figshare. Dryad charges a fee that is waived for US NSF-funded authors.

Computing Center: High-Performance Computers (HPC) have the infrastructure to support research using models and simulations, which may be involved in generating and/or analyzing high volumes of data. The operations team at the center may have recommendations for data management, storage, and preservation.

National Repositories: Some countries require authors to use their National Repositories for data and/or software preservation.

During Peer Review

Data: Your data must be available for peer review of your manuscript. Here are options to ensure confidential access to your data.

1. Preserve your data in a repository and make it available for peer review. Depending on the repository, this process can be done in a couple of ways:

   a. Provide a temporary private link ("share link") in the last sentence of the Open Research section of your paper. This link will not be present in your published

paper as it is not a persistent link. This approach allows your data to remain private until acceptance of the manuscript for publication.

    b.   Provide the persistent identifier (e.g., DOI) for your data; i.e., when your dataset has completed the repository submission process and is now publicly available.

2. Include your data in the supplementary information of your paper, only <u>for the purpose of peer review</u>. The supplement is not a repository and can only be used to support the peer review process. You must still submit your data to a repository prior to paper acceptance.

<u>Software</u>: For papers where software is central to the research, your software, or workflow, must be available, if the peer reviewer needs access. The options for providing or limiting access to your software are the same as for data.

Peer reviewers, with support from AGU staff, will ensure the link to the data and software resolves properly to a community-accepted, trusted repository and includes the data and software necessary to evaluate your research.

Paper Acceptance

1. <u>Data</u>: Relevant data and model results should be accessible at the time your paper is accepted. Note the possibility that in unusual cases the repository policy may not allow your data to be published until your paper is published. If that is the case, AGU will accept that your data will be made available after your paper is published. Please coordinate with the repository to ensure the availability of your data.

2. <u>Software</u>: For papers where software is central to your research, your software should be accessible at the time your paper is accepted.

Availability Statements and Template Examples

The Availability Statement is a narrative that indicates to the reader where and how to directly access your data and software and provides any information on licenses and restrictions. An Availability Statement should contain an in-text citation, licensing information (e.g., CC-BY 4.0, MIT), and access restrictions (e.g., authentication required) (here is an [example from JAMES](#)). Statements to the effect of "data available from authors" are not acceptable. Also, statements to the effect of "data available from http://nasa.gov" are not acceptable since high-level website references do not meet the [AGU's Data and](#)

[Software for Authors](#) requirements. Provide details of the specific locations, data product names, variable names, time ranges, spatial locations, and any other search criteria needed to Find and Access the data used/ generated in the paper (including those represented in figures and tables).

Data Availability Statement:

1. **Data archived in a repository:** Datasets for this research are available in these in-text data citation references: Smith et al. (2019), [with this license, and these access restrictions if any], Jones et al. (2017) [with this license, and these access restrictions if any]. Such datasets must be findable and accessible (e.g. via URLs).

2. **Data published in the literature:** Datasets for this research are included in this paper (and its supplementary information files): [citation for paper] or point to where the references are compiled. Such datasets *must be findable and accessible* (e.g. via URLs). For example:

   Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. A. Parsons and M. J. Brodzik. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5060/D4MW2F23. Accessed 2008-05-14. *Reproduced from ESIP

Citations for data in regular publications where data are not findable or accessible, i.e. not available are NOT acceptable. For a made-up example:

   Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002. CLPX-Ground: ISA snow depth transects and related measurements, J. Ice., vol 1 (2), pp. 3-9. (journal is subscription only, and data are not available in the article or supplement information or is in a proprietary format that is no longer readable).

3. **Technical reports publishing the description of a dataset and its preparation, e.g., a data paper:** Datasets for this research are described in this paper: [citation for paper, with this license, and these access restrictions if any]. Such datasets must be findable and accessible (e.g. via URLs).

4. **Theoretical papers, or most review papers:** Data were not used, nor created for this research.

5. **Data not publicly available, but available to researchers with appropriate credentials:** Data for this research are not publicly available due to [Fill in reasons]. Data are stored in this in-text data citation reference: Smith et al. (2019), [with this

license, and these access restrictions if any].

6. **Data that are restricted by commercial, industry, patent, government policies, regulations or laws:** Data supporting this research are available in [cite in-text data citation reference from third party source], with [these restrictions that include information concerning required NDA, licensing, agreements], and are not accessible to the public or research community. [Provide a process for how other researchers can gain access.] NOTE: If your data are in this category, the editors will determine if this statement meets the AGU data guidelines sufficiently.

Software Availability Statement:
The Availability Statement should contain a citation, licensing information, access restrictions, and a link to the development platform (e.g. GitHub). Note that "git/GitHub/GitLab" are not acceptable software repositories because they are not archival (see AGU's Data and Software for Authors).

1. **Software archived in a repository:** Software for this research is available in these in-text data citation references: Smith et al. (2019), [with this license, and these access restrictions if any], Jones et al. (2017) [with this license, and these access restrictions if any]. Such software must be findable and accessible (e.g. via URLs).

2. **Software published in the literature as supplementary information:** Software for this research is included in this paper (and its supplementary information files): [citation for paper] or point to where the references are compiled. Such software must be findable and accessible (e.g. via URLs).

3. **Software not publicly available, but available to researchers with appropriate credentials:** Software for this research is not publicly available due to [Fill in reasons]. Software is stored in this in-text citation reference: Smith et al. (2019), [with this license, and these access restrictions if any].

4. **Software that is restricted by commercial, industry, patent, government policies, regulations or laws:** Software supporting this research is available in [cite in-text citation reference from third-party source], with [these restrictions that include information concerning required NDA, licensing, agreements], and is not accessible to the public or research community. [Provide a process for how other researchers can gain access.] NOTE: If your software is in this category, the editors will determine if this statement meets the AGU guidelines sufficiently.

5. **Theoretical papers, or most review papers:** Software (other than for typesetting) was not used for this research.