

Steps Toward
**Incentivizing data and software
sharing**

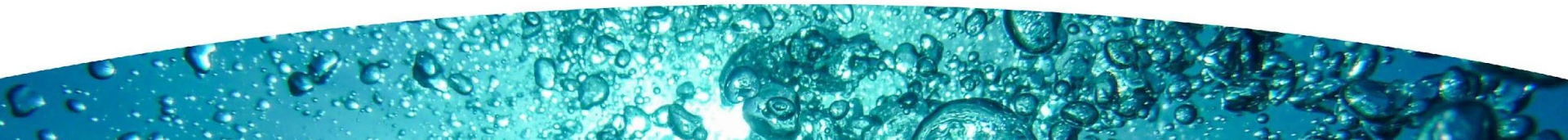
EarthCube Model Data RCN
25 July 2022

Chris Erdmann
Assistant Director, Data Leadership
American Geophysical Union
0000-0003-2554-180X
@libcce | cerdmann@agu.org

AGU's position statement on data affirms that

“Earth and space science data are a **world heritage**, and an essential part of the science ecosystem. All players in the science ecosystem—researchers, repositories, publishers, funders, institutions, etc.—should work to **ensure that relevant scientific evidence is processed, shared, and used ethically, and is available, preserved, documented, and fairly credited.**”

https://www.agu.org/Share-and-Advocate/Share/Polycymakers/Position-Statements/Position_Data



Researchers, Journals, Data Repositories
NSF Grant 2025364

Accelerating Open and FAIR Data Practices Across the Earth, Space, and Environmental Sciences: A Pilot with the NSF to Support Public Access to Research Data (AGU NSF PAR 2.0)



Partners: Dryad, CHORUS, ESIP, Wiley (In-Kind)

2-year project aimed at implementing FAIR data practices across the Earth, space, and environmental sciences such that, by the end of the project:

- Data citations for data funded by NSF grants are captured in the NSF Public Access Repository (PAR 2.0)
- Knowledge of leading practices and workflows around data citation are well known across the AGU community.

AGU Data & Software Sharing Guidance

What is covered:

- What data needs to be available?
- Repository Selection
- Availability Statement
- Data & Software Citation
- Citation Formatter
- Models & Simulations
- Journal Specific Guidance
- International Geo Sample Numbers
- Data Help Desk

AGU ADVANCING EARTH AND SPACE SCIENCE

JOIN RENEW GIVE LOGIN Q

Data & Software for Authors

WHAT IS NEEDED?

AGU requires that the underlying data needed to understand, evaluate, and build upon the reported research be available at the time of peer review and publication. Additionally, authors should make available software that has a significant impact on the research. This entails:

1. Depositing the data and software in a trusted repository, as appropriate, and preferably with a DOI
2. Including an [Availability Statement](#), as a separate paragraph in the Open Research section explaining to the reader where and how to access the data and software
3. And including [citation\(s\)](#) to the deposited data and software, in the Reference Section.

Click on the headings below for detailed information on:

- [Models & Simulations](#)
- [Journal-Specific Data Guidance](#)
- [International Geo Sample Numbers](#)

Most of your questions regarding data and software should be answered by the resources below. Just in case, if you still have questions, you can contact DataHelp@agu.org.

WHAT DATA NEEDS TO BE AVAILABLE?

Primary and processed data used for your research should be preserved and made available. Generally, the underlying data are considered to be the types of data usually preserved in domain repositories for each discipline. These may include raw data, but are usually the processed or refined data that support and lead to the described results and allow other readers to assess your conclusions and build off your work.



In your paper, cite these data, as well as any data you used from other sources, and include information about access to the data in the availability statement. For model or simulation data, follow [journal specific guidance](#) on prioritizing preserved output; in general, availability of software is most important.

Very large data (greater than 1 terabyte or TB) can be a challenge to preserve as there often fees and additional resources required. One option to consider, institutions often offer solutions for data preservation and compliance. Again, refer to the [journal specific guidance](#) for more information or email DataHelp@agu.org.

<https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-and-Software-for-Authors>

Availability Statement/Citation Example

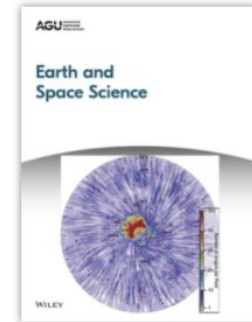
Earth and Space Science

Research Article | [Open Access](#) |   

Performance Assessment of Geophysical Instrumentation Through the Automated Analysis of Power Spectral Density Estimates

M. R. Koymans , J. Domingo Ballesta, E. Ruigrok, R. Sleeman, L. Trani, L. G. Evers

First published: 22 July 2021 | <https://doi.org/10.1029/2021EA001675>



[Volume 8, Issue 9](#)

September 2021

e2021EA001675

<https://doi.org/10.1029/2021EA001675>

Availability Statement/Citation Example (Cont.)

Example: Koymans, M. R., Domingo Ballesta, J., Ruigrok, E., Sleeman, R., Trani, L., & Evers, L. G. (2021). Performance Assessment of Geophysical Instrumentation Through the Automated Analysis of Power Spectral Density Estimates. In *Earth and Space Science* (Vol. 8, Issue 9). American Geophysical Union (AGU). <https://doi.org/10.1029/2021ea001675>

From Open Research (Availability Statement) Section: Maps were created through PyGMT (prerelease) (Uieda et al., 2021) using Generic Mapping Tools (GMT) version 6 (Wessel et al., 2019a, 2019b) licensed under LGPL version 3 or later, available at <https://www.genericmapping-tools.org/>.

References:

Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., & Tian, D. (2019a). The generic mapping tools version 6 [Software]. Zenodo. (Funded by US National Science Foundation grants OCE-1558403 and EAR-1829371). <https://doi.org/10.5281/zenodo.3407866>

Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., & Tian, D. (2019b). The generic mapping tools version 6. *Geochemistry, Geophysics, Geosystems*, 20(11), 5556– 5564. <https://doi.org/10.1029/2019gc008515>

Soon to be submitted for publication...

Journal Production Guidance for Software and Data Citations - Draft

xx May 2022 - Draft Review

Journal Production Guidance for Software and Data Citations

Authors: Shelley Stall, ...

Abstract

Software and data citation are emerging best practices in scientific communication. This article provides structured guidance to the academic publishing community on how to implement software and data citation in publishing workflows. These best practices support the verifiability and reproducibility of scientific results, sharing and reuse of valuable data, software tools, and resources, and credit to the originators of the data and software. They provide a basis for making both data and software FAIR (Findable, Accessible, Interoperable, and Reusable). Data citation is becoming increasingly well-established. With the current intensive use of software, including specialized tools and models for scientific research problems, the research community has begun to recognize that software, as a key research resource, requires the same level of transparency, accessibility, and disclosure as data. Software and data that support scientific results should be preserved and shared in scientific repositories for discovery, transparency, and used by other researchers. This can be achieved by citing these products in the references section of papers and effectively associating them.

AGU Journals - Data/Software Citations (2021)

JOURNAL	PAPER AVAILABILITIES INTEXT CITATION COUNT	TOTAL PAPERS	PERCENTAGE
JOURNAL OF GEOPHYSICAL RESEARCH: PLANETS	209	274	76.28%
JOURNAL OF GEOPHYSICAL RESEARCH: SOLID EARTH	341	809	42.15%
GEOCHEMISTRY, GEOPHYSICS, GEOSYSTEMS	132	332	39.76%
TECTONICS	77	197	39.09%
GLOBAL BIOGEOCHEMICAL CYCLES	50	140	35.71%
JOURNAL OF GEOPHYSICAL RESEARCH: EARTH SURFACE	62	191	32.46%
REVIEWS OF GEOPHYSICS	7	23	30.43%
JOURNAL OF GEOPHYSICAL RESEARCH: BIOGEOSCIENCES	96	323	29.72%
WATER RESOURCES RESEARCH	201	716	28.07%
GEOHEALTH	26	93	27.96%
JOURNAL OF ADVANCES IN MODELING EARTH SYSTEMS	60	215	27.91%
PALEOCEANOGRAPHY AND PALEOCLIMATOLOGY	132	483	27.33%
EARTH'S FUTURE	49	181	27.07%
GEOPHYSICAL RESEARCH LETTERS	494	1856	26.62%
EARTH AND SPACE SCIENCE	86	325	26.46%
SPACE WEATHER	74	340	21.76%
JOURNAL OF GEOPHYSICAL RESEARCH: OCEANS	112	529	21.17%
AGU ADVANCES	16	76	21.05%
JOURNAL OF GEOPHYSICAL RESEARCH: ATMOSPHERES	146	843	17.32%
JOURNAL OF GEOPHYSICAL RESEARCH: SPACE PHYSICS	135	788	17.13%
RADIO SCIENCE	15	108	13.89%
PERSPECTIVES OF EARTH AND SPACE SCIENTISTS	0	9	0.00%
Year 2021 counts	2520	8851	28.47%



Domain-Discipline Repositories Useful to AGU Journals

OCTOBER 24, 2021

Domain-Discipline Repositories Useful to AGU Journals

The data that supports the research reported in your paper must be deposited in a community-accepted, trusted preservation repository. Additionally, authors should make available software that has a significant impact on the research. A repository that specializes in domain-discipline specific data and software will maximize the probability that the deposited data and software will be findable, accessible, interoperable and reusable (FAIR). Repositories that use persistent identifier links (e.g. DOI or digital object identifier over URLs (and not to the home page) are recommended. Note, an English language translation is necessary in order for the data/software to be accessible to the wider community. Domain-discipline repositories useful to AGU journals below may also be at different stages in supporting the FAIR principles. For any additional domain-discipline repositories recommendations, contact datahelp@agu.org or [submit a GitHub issue/pull request](#). Otherwise, look to your [institutional repository](#), your computing center, a [general repository](#) (e.g., [Zenodo](#), [Dryad](#), [figshare](#)), or search for a repository using [re3data](#), [OpenAire](#), or [DataOne](#). Consult [Data and Software for Authors](#) and [Data and Software Sharing Guidance for Authors Submitting to AGU Journals](#) for more in-depth guidance.

<https://data.agu.org/resources/useful-domain-repositories>

From the Help Desk

- Government Sites, Similar - Technical, Permissions
- Firewalls, Authentication - Openness, Availability, Anonymity
- Supplemental Information - Tradition, Peer Review
- FTP, Directories, Storage - Institutional, Compliant Solution
- Curation, Deposit Workflows - Service, Publication Workflows
- Web Sharing, Dev Platforms - Citation Information
- Databases / Dynamic Services - Direct Access, Linking
- Available Upon Request - Culture
- Citation Nothingness (Paper not the Data) - Culture
- Website Home (Parachuting) - Laziness
- English Language - Language Diversity, Translation
- Many Data Links/Citations - Tables, Supplements (See [Data Citation Community of Practice](#))
- ...

Preserving Large Data!



Preserving very large data is a challenge. Spoilers, there are no easy answers!

OCTOBER 01, 2021

When it comes to large datasets, we are often asked by authors and editors how they should preserve the data. These questions come via datahelp@agu.org and our [data and software guidance](#) discussions. Spoilers, there are no easy answers, yet! Here we offer our experience, share the current limitations, and the approaches we recommend with what is possible right now.

AGU requires that primary and processed data used for your research should be preserved and made available. This can range from observational data to the data used to generate your figures. The raw data may be needed, but usually, the processed or refined data that support and lead to the described results and allow other readers to assess your conclusions and build off your work should be preserved.

For data that is large, over 1 Terabyte (TB), authors run into the challenge of finding a suitable repository. Many repositories have file size limitations but also costs associated with deposits over certain limits. This [generalist repository comparison chart](#) provides an overview of the limitations. Discipline-specific and institutional repositories are often a place to turn to for assistance with preserving large data but they also have limitations and potential costs. This emphasizes the importance of avoiding surprises at the time of publication by:

<https://data.agu.org/2021/10/01/challenges-preserving-very-large-data.html>

Models & Simulations

RCN -Determining Best Practices for Preservation and Replicability of Model Data

Doug Schuster, NCAR
Matt Mayernik, NCAR
Gretchen Mullendore, NCAR/U. North Dakota



NCAR | NATIONAL CENTER FOR
ATMOSPHERIC RESEARCH

<https://modeldatarcn.github.io/>

NSF Awards #1929773, #1929757



Mullendore, Gretchen, Schuster, Doug, Mayernik, Matthew, & Griffies, Stephen M. (2021, May). COPDESS Workshop: Rubric for Models and Model Data – Best Practices for Preservation and Replicability. Zenodo.

<http://doi.org/10.5281/zenodo.4890691>

Guidelines for Research Primarily Based on Numerical Models or Theory

2. 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

1. 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).
2. 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.

1. 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.

1. 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.

1. 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.
2. 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.
3. ACCEPTABLE OPTION: Provide the necessary information for transparency and traceability of the analysis using your community standards or guidance.

5. Model Output Data.

1. 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.

<https://data.agu.org/resources/agu-data-software-sharing-guidance#guideline>

Notebooks Guidance for Authors ⇒ Notebooks Now!



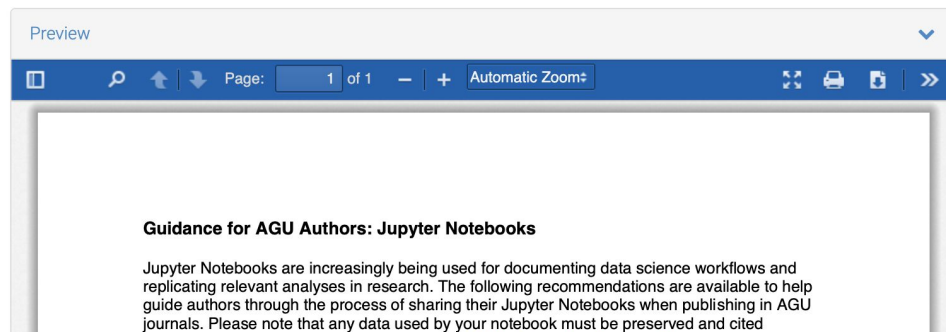
May 20, 2021

Other Open Access

Guidance for AGU Authors - Jupyter Notebooks

Erdmann, Christopher; Stall, Shelley

Jupyter Notebooks are increasingly being used for documenting data science workflows and replicating relevant analyses in research. The following recommendations are available to help guide authors through the process of sharing their Jupyter Notebooks when publishing in AGU journals. Please note that any data used by your notebook must be preserved and cited separately from the notebook to comply with [AGU's data and software guidance](#).



Erdmann, Christopher, Stall, Shelley, Gentemann, Chelle, Holdgraf, Chris, Fernandes, Filipe P. A., & Gehlen, Karsten Peters-von. (2021, May 20). Guidance for AGU Authors - Jupyter Notebooks. Zenodo.

<http://doi.org/10.5281/zenodo.4910038>

Search/Indexing/Filtering/Aggregation



Advancing FAIR in the US

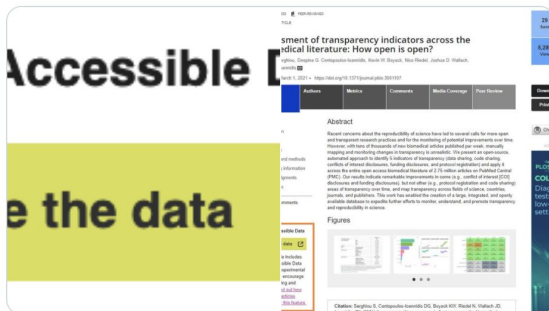
Our Aim is to connect FAIR stakeholders and foster a community where FAIR approaches can be shared, discussed, and advanced collaboratively.

[Learn More](#)



iain hrynaskiewicz
@iainh_z

We've launched a new experimental feature @PLOS journals to see if adding prominent links to research data stored in repositories on article pages increases data use and/or incentivises use of data repositories. With thanks to support from @wellcometrust thepliosblog.plos.org/2022/03/access...



Figshare and 2 others

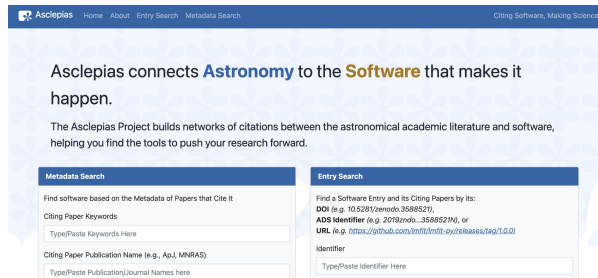
4:54 AM · Mar 30, 2022 · Twitter Web App

54 Retweets 10 Quote Tweets 129 Likes

<https://www.gofair.us/>

https://twitter.com/iainh_z/status/1509091657131638792

<https://asclepias.aas.org/>



Open Science!



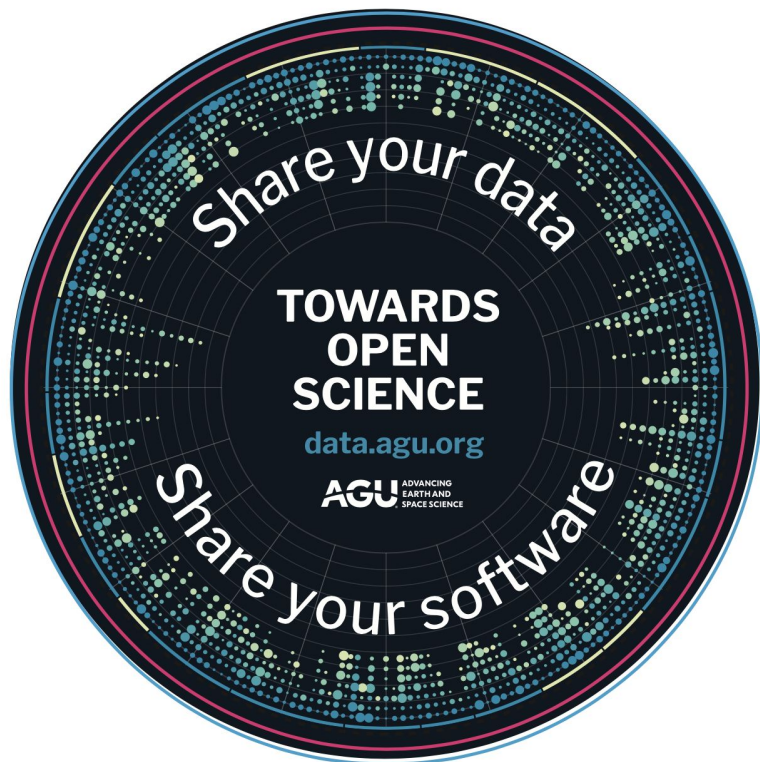
NATIONAL
ACADEMIES *Sciences
Engineering
Medicine*

[Transform to Open Science \(TOPS\) | Science Mission Directorate \(nasa.gov\)](#)

[UNESCO Recommendation on Open Science](#)

[Developing a Toolkit for Fostering Open Science Practices A Workshop | National Academies](#)

Thank you.



Chris Erdmann

Director, Data Leadership

American Geophysical Union

0000-0003-2554-180X

@libcce | cerdmann@agu.org

