Do We Have a Reproducibility Crisis: How Available is Data and Code Across Journals in Artificial Intelligence and Earth Sciences?

Erin A. Jones[a] , Brandon McClung[a] , Hadi Fawad[c] , Amy McGovern[a,b,c]

[a] *School of Meteorology, University of Oklahoma, Norman OK USA*

[b] *NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), USA*

[c] *School of Computer Science, University of Oklahoma, Norman OK USA*

Corresponding author: amcgovern@ou.edu

ABSTRACT: As the use of artificial intelligence (AI) has grown exponentially across a wide variety of science applications, it has become clear that it is critical to share data and code to facilitate reproducibility. AMS recently adopted the requirement that all papers include a data availability statement. However, there is no requirement to ensure that the data and code are openly accessible during and after publication. Studies show that without this requirement, data is openly available in about one third of journal articles. In this work, we surveyed two AMS journals, AIES and MWR, and two non-AMS journals, considering the following research questions. First, to what extent are the data and code stated to be available in AIES journal articles? Second, how do these results compared to articles in 1) MWR, an AMS journal without a primary focus on AI; 2) a non-AMS journal with a data availability statement requirement focused on AI but not Earth sciences; and 3) a non-AMS journal focused on AI in Earth sciences without a data availability statement requirement? Third, for the papers which claim to have openly accessible data and code, can readers easily access the data and code? Finally, what are the justifications that are provided for articles that have a data availability statement but do not provide open access to their data or code?

SIGNIFICANCE STATEMENT: Making code and data available to future researchers is critical for research reproducibility. Despite this, if it is not required, authors share their code and data only about one third of the time. We show that even with the new AMS journal requirement to include a data availability statement, the actual availability is limited. This issue is important to address for future research, and especially with the growing research in AI. If data and models are made easily available, people can innovate on these models in a more equitable manner.

1. Introduction

There has been a recent rapid acceleration of growth of the use of artificial intelligence (AI)—both as a tool in Earth science research as well as in society as a whole (e.g., Haupt et al. 2022; Stall et al. 2023; Maslej et al. 2024). AI tools increasingly have complex architectures, which may be a barrier for scientific innovation and reproducibility (e.g., Pineau et al. 2021; Liesenfeld and Dingemanse 2024). These tools also rely on copious amounts of training data, which rely on the producers of the AI to have sourced ethically and without bias (e.g., McGovern et al. 2024; Wirz et al. 2024). Increased transparency, however, may be obtained through the documentation and open sharing of training data, pre-processing and model code, and any associated metadata. The availability of shared resources expedites collaborative efforts, which is essential for tackling multifaceted challenges with global societal impacts (e.g., Stall et al. 2023).

Recently, AMS journals adopted a policy with the expectation that a data availability statement (DAS) accompanies every published article[1]. AMS is not alone in this expectation. Internationally, science is becoming more open (e.g., Grant and Hrynaszkiewicz 2018; Graf et al. 2020; UNESCO 2021; Bertram et al. 2023). Several ethical guidelines have been developed to help scientists navigate making their research more open (e.g., Goodman et al. 2014; Fecher and Friesike 2014). AMS, specifically, cites the FAIR (Findable, Accessible, Interoperable, and Reusable) Guiding Principles (Wilkinson et al. 2016) in their commitment to open data[2]. These principles suggest not only that datasets and code are easily available, but also that they are supplemented with appropriate documentation and metadata so that any research conducted using them can be reproduced.

[1] https://www.ametsoc.org/index.cfm/ams/publications/ethical-guidelines-and-ams-policies/data-and-software-policy-guidelines-for-ams-publications/

[2] https://www.ametsoc.org/index.cfm/ams/about-ams/ams-statements/statements-of-the-ams-in-force/full-open-and-timely-access-to-data

Although a DAS is required by AMS policy, fully open data or code is only recommended and it is up to the individual reviewers to enforce that the data URLS provided are valid. Without a specific requirement to make data openly available, studies have found that only about a third to a half of published works with required DASs have open data (Grant and Hrynaszkiewicz 2018; McGuinness and Sheppard 2021; Tedersoo et al. 2021; Campbell and Mu 2023).

Given the rapid advances being made in AI, including within the atmospheric and related science community, we will focus our study on DASs from four journals in the fields of AI and/or Earth Science. First, we will examine DASs from the AMS journal, Artificial Intelligence for the Earth Systems (AIES) to determine the level of data and code availability provided. To compare AMS journals with varying research foci, we will also examine DASs from Monthly Weather Review (MWR), which does not have a primary focus on AI applications. Additionally, we will examine two non-AMS journals: Artificial Intelligence in Geosciences (AI in Geo.) and Artificial Intelligence (AIJ). Similar to AIES, AI in Geo. also has a focus on AI applications in Earth Sciences. However, it does not have a DAS requirement, allowing us to examine the impact of such a requirement. AIJ has a similar DAS requirement to AMS journals. Additionally, AIJ has a primary focus on advancements of AI without concentrating on Earth Science applications, allowing for further comparisons to be made across primary disciplines.

2. Data and Methods

For each journal, the years and number of articles examined are given in Table 1. Given the relatively limited repertoire of AIES and in AI in Geo., all articles published before 15 April 2024 and their associated DASs were examined. MWR and AIJ each have a much larger yearly and total number of articles. Therefore, only a sample of articles were examined for each journal.

For each article, we collected and recorded the metadata and general information about topic of each article as well as categorized the information about data and/or code (DaCo, hereafter) availability in the DAS, if one was provided. Data availability was categorized as follows: 1) all data openly available; 2) at least some data openly available; 3) data available upon request; 4) no data produced; 5) data not available; 6) no DAS provided. All DASs were subjectively categorized. For example, if it was not clearly stated that some data were not openly available, the DAS was likely placed in category 1. Code was categorized similarly, except the "available upon request"

TABLE 1. Description of journals and number of articles for each journal belonging to each DAS category.

	AIES	AI in Geo	MWR	AIJ
Publisher	AMS Journals	KeAi Publishing	AMS Journals	Elsevier
Online Distribution Platform	AMS Journals	ScienceDirect	AMS Journals	ScienceDirect
Years analyzed	2022-2024	2020-2024	2023	2023-2024
Total articles examined	107	72	54	55
Articles with DASs	107	21	53	55
All data available	76	12	25	12
Some data available	14	1	10	0
Data available upon request	4	5	11	9
No data produced	8	1	0	30
No data available	5	2	7	4
No DAS	0	51	1	0
All code available	56	3	13	15
Some code available	2	0	1	0
No code produced	8	1	0	30
No code available	41	17	39	10
Articles without broken links	84	11	34	12
Articles with broken links	10	3	8	3

category was not separately analyzed. If the DAS claimed DaCo was available, we further recorded its accessibility, such as if any links in the DAS were broken or led to unrelated websites. Finally, if the any DaCo was unavailable, we noted any stated justification.

3. To what extent is data openly shared?

In AIES, all 107 articles examined were submitted after the AMS mandate that every article contain a DAS. Of those articles, 84.1% claimed to make some or all of the data used and produced by the study openly available (Fig. 1a). The DASs of an additional 7.5% of articles claim that their associated work did not utilize any datasets or produce any data. These articles were largely “Perspectives,” “Review,” or “Lessons Learned” article types. Only 3.7% and 4.7% of DASs state that data are available upon request or not available respectively. The proportion of articles with data available is larger than the approximately third to half of all articles found in prior literature (Grant and Hrynaszkiewicz 2018; McGuinness and Sheppard 2021; Tedersoo et al. 2021; Campbell and Mu 2023).

We also examined all articles published in AI in Geo. as another journal with a focus on AI in Earth Science, though not frequently atmospheric science. As AI in Geo. does not require a

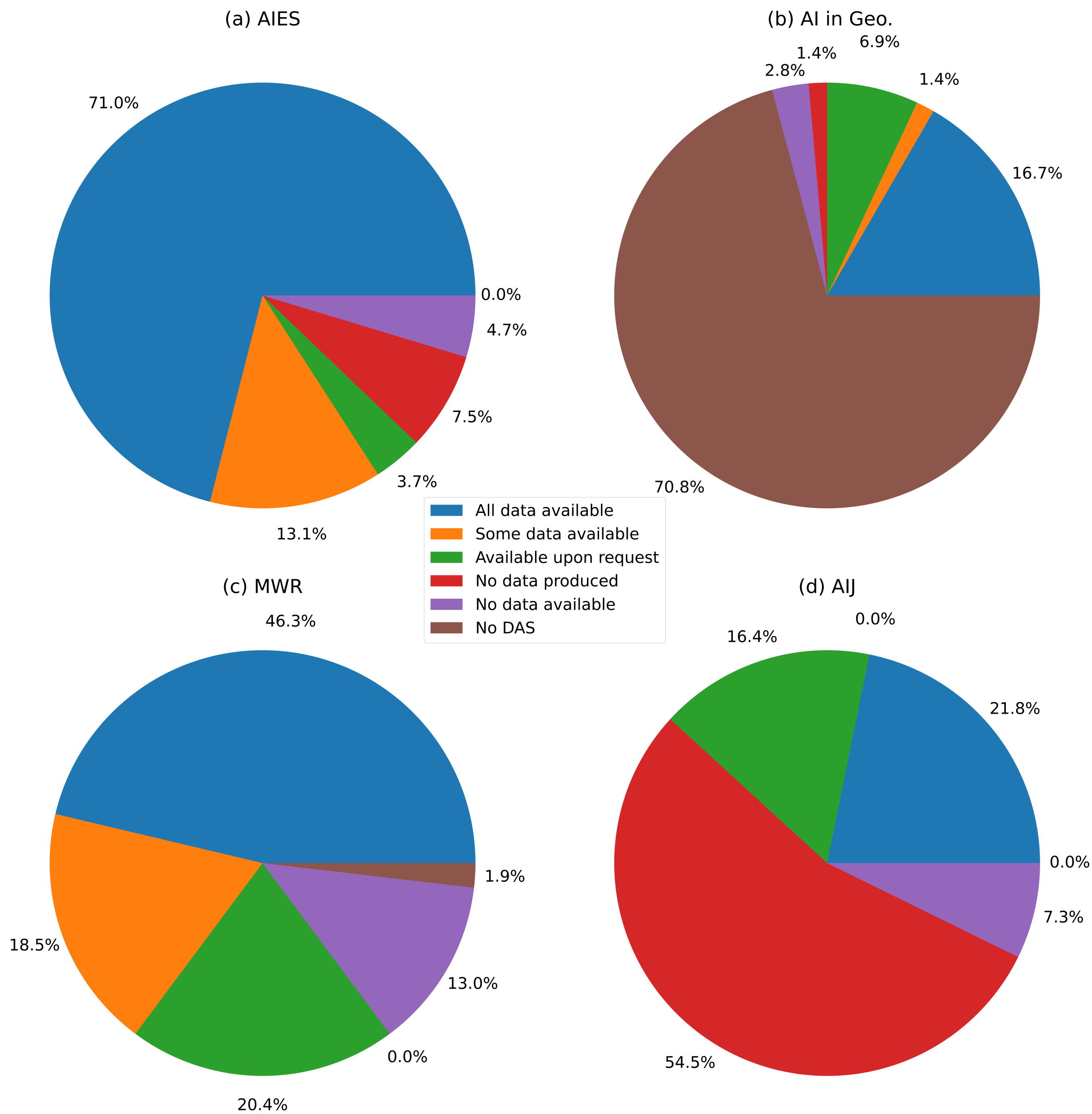

Fig. 1. Proportion of open availability of data for (a) AIES, (b) AI in Geo., (c) MWR, and (d) AIJ.

DAS to be submitted with the publication, 70.8% of articles did not include one (Fig. 1b). Of the remaining articles examined, however, 14 out of 21 had at least some data openly available or no data produced. Only 2 DASs did not make any data available. Additionally, although there was no DAS requirement at any point, the percentage of articles with a DAS in increased steadily from 2021 through 2023 and was on track to further increase in 2024 (not shown).

Compared to AIES, the 54-article sample chosen for analysis from MWR has a slightly smaller percentage of journals with some or all data openly available at 64.8% (Fig. 1c). Around 20.4% of DASs stated that data would be made available upon request; 14.9% of articles did not make data available, including one article that did not include a DAS as it was first submitted prior to the enforcement of the DAS requirement. The percentage of non-available data is substantially larger for MWR compared to AIES. MWR often publishes research involving large numerical modelling or data assimilation experiments, where dataset size may be unfeasible to store and maintain openly. Although authors should strive for as much open availability as possible, following guidance for the publication of model data such as in Schuster et al. (2023), this reasoning may explain some of increase in non-availability compared to AIES.

Slightly over half of our AIJ articles examined were pure theory and review papers, so no data were produced for these articles (Fig. 1d). Of the remaining 25 articles, 12 DASs made all data openly available; 9 DASs made data available upon request; and 4 DASs did not describe openly available data.

When DASs were present and data were produced for the study, at least some data were stated to be openly available in more than half of the DASs examined for each journal and more than three quarters as a total between all journals. This result is a larger estimate compared to prior literature (Grant and Hrynaszkiewicz 2018; McGuinness and Sheppard 2021; Tedersoo et al. 2021; Campbell and Mu 2023). Though the specific reason for this discrepancy is beyond the scope of this article, these results are potentially indicating a cultural shift in the perceived value of open data.

4. Is code openly shared to the same extent as data?

Just over half of the DASs in AIES provided links to openly available code (Fig. 2a). AI in Geo. and MWR similarly have substantially fewer DASs providing code than data at around 5% and around 25% of all articles examined respectively (Fig. 2b,c). In AI applications, providing code for the model along with any training pipelines and post-processing steps are just as essential as providing training datasets for scientific reproducibility and transparency (e.g., Liesenfeld and Dingemanse 2024). Similarly, open access to numerical model, data assimilation, post-processing, and/or statistical verification code is also as important as data used or produced. Although, in their

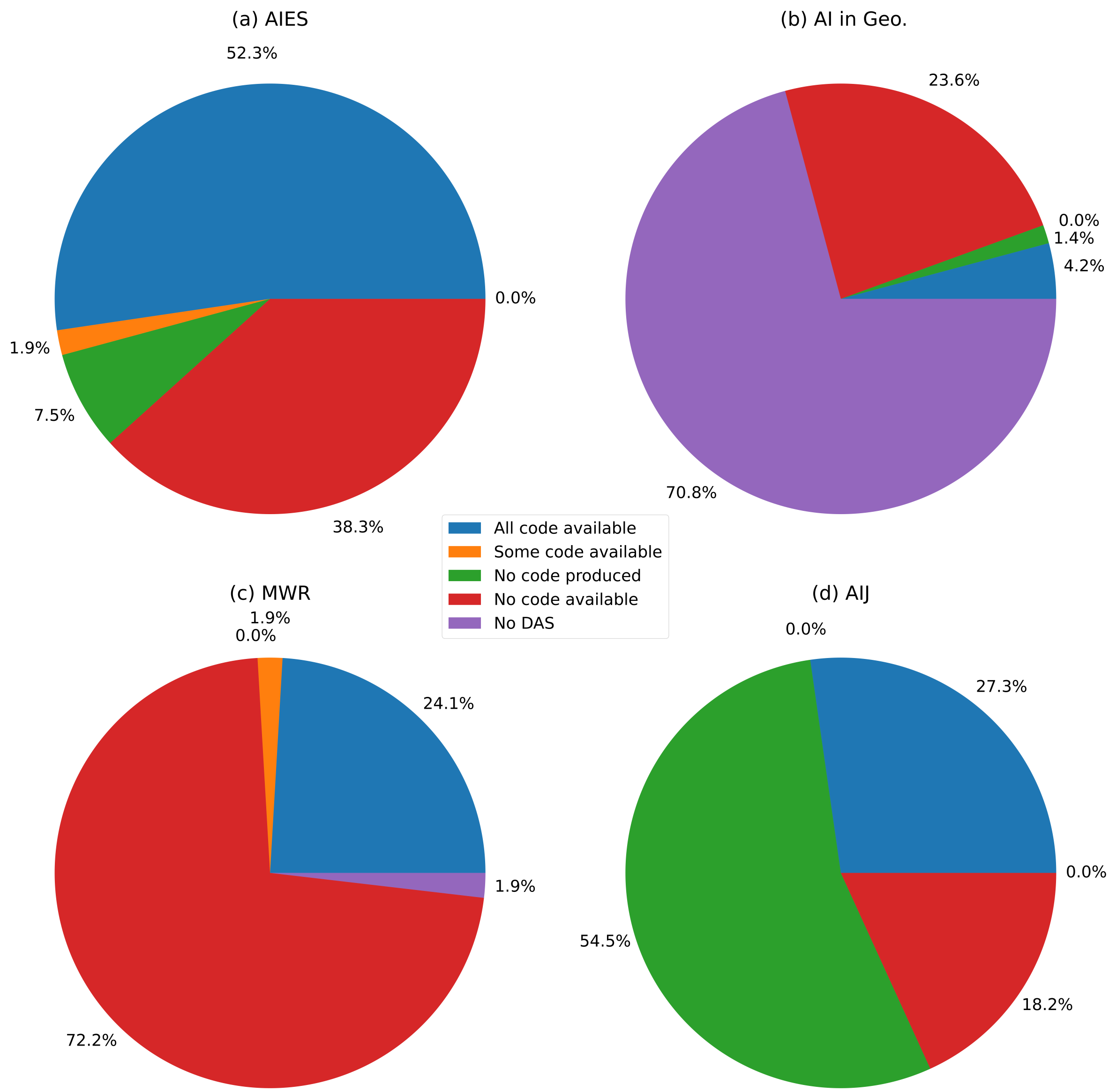

Fig. 2. As in Fig. 1 but for open availability of code.

guidelines, AMS indicates that any software used or produced for the articles published should have a reference and a link provided, authors may not consider providing software as essential in a "data" availability statement. Additionally, there may not be consistent enforcement of this policy between AMS journals.

Contrasting from the three journals focused on Earth Science applications, AIJ has a larger proportion of articles that provide open code—at 15 out of the 25 articles where DaCo is produced—compared to open data—at 12 out of 25 articles. This difference may indicate some contrasting culture in what is meant by "open" and what is most valuable to share in research with a focus on Earth Science compared to solely AI.

5. Is "open" data or code actually accessible?

For Figs. 1-2, we examined the DASs for stated availability of data and code. Even if data and code were stated to be available, they may not be easily accessible. In an examination of research produced from a single university, Briney (2024) noted that approximately 5% of links to data were no longer available, making it difficult or impossible for readers to access supposedly open data. The study also noted that the percentages of unavailable links increased with time from the initial publication of the data.

In our examination of the 216 DASs in this study, we also determined how many DASs included links to DaCo repositories. We verified each link provided to determine if it directed the reader to the repository or if the link was 'broken' and did not lead to a currently established webpage. We determined that 165 DASs included at least one link to a DaCo repository of which approximately 15% contained at least one broken link (Table 1). Out of the four journals AIES had the smallest proportion of broken links at approximately 11% of articles.

In addition, throughout our examination of links, we found that the web page to which a link directed often did not lead directly to a DaCo repository. Such links frequently led to project or agency home pages, where the DaCo was not easily accessible or occasionally not findable at all. While directing a reader to these home pages may be useful to establish context about a code or dataset, these types of links should not be provided in substitution for direct links.

It may, therefore, be prudent for AMS and other journal publishers to create policies to ensure during the review process that any links that claim to point readers to code or dataset actually send readers as directly as possible. Additionally, a periodic examination for broken links in all DASs would help to ensure that data and code remain open as intended for all readers.

6. What are common justifications for not having openly available data or code?

For every DAS that partially or fully did not provide open availability to DaCo, we recorded whether a justification was given for why there was unavailability. If a justification was given, we

FIG. 3. For all journals, a word cloud for all justifications as to why data is not openly available. The word cloud is generated by `https://www.jasondavies.com/wordcloud/`.

noted a summary of the reasoning. Fig. 3 shows a word cloud using the summarized justifications from all four journals. Larger words within the cloud are associated with greater frequency.

The most frequent justifications could generally be sorted into five categories. First, datasets used for the article were too large to be published openly. Second, there were issues with licensing, or the DaCo was proprietary. Third, the DaCo are not made openly available by the authors of the article but can be obtained from other entities, such as by contacting a specific government agency or individuals who are not co-authors on the article. Fourth, the data contain sensitive information, such as human subject research, controlled unclassified information, or information relevant to national security. Fifth, the DaCo is not currently available, but they will be made openly available once funding for the associated project concludes. While this list is incomplete for the full set of reasons that may be utilized by an author when deciding not to provide DaCo, each category of justification provides insight into what challenges need to be overcome in order to provide fully open DaCo with each.

7. What recommendations do we suggest to further promote open science?

a. For authors?

Regardless of the requirements of the journal, we recommend that authors provide a DAS with their manuscript. DASs aid in allowing for reproducibility and advancement of science as well as enhancing trust with readers. Trust is especially important in rapidly advancing and broad-impact fields, such as AI. The DAS should provide links to repositories where readers can access any DaCo used or produced for the article. Within the repositories, authors should provide metadata and documentation so that the DaCo is interoperable and reusable for further research purposes (e.g., Edwards et al. 2011). The repositories should be maintained so that the link associated with them stays active. Preferably, a Digital Object Identifier (DOI) should be obtained as typically, these have greater digital permanence than a general URL (Briney 2024). Authors should check the repositories with their DaCo periodically. If any links change, they should contact any journals publishing articles containing such links. If there is some limitation to the open publication of any DaCo, authors should still provide any DaCo they are able. Additionally, these authors should clearly provide justification within their DAS for what is not available and provide clear directions for obtaining any DaCo that may be accessed by some means (e.g., by sending a request to a government agency) but not open to all.

b. For research journal publishers?

The onus of open science should not be solely on the authors. Graf et al. (2020) showed that the number of articles including a DAS increase with the mandate of a DAS from the journal. We recommend for journals, such as AI in Geo., who do not currently have a DAS mandate, to make such a policy a priority. We encourage editors and peer reviewers to examine any DaCo repositories provided to ensure that direct links are given to DaCo and sufficient metadata and documentation is given with the DaCo. We support journals exploring a system which would remind corresponding authors to periodically check for broken links within their articles and give a simple means to update such links. In a rapidly changing environment where AI is increasingly being leveraged in the sciences, it is imperative for journals to evolve their practices to ensure transparency and accessibility. Adapting to these advancements will not only uphold scientific integrity but also set a new standard for future research publications. Finally, we recommend that journals provide authors with a clear set of guidelines for which justifications, if any, are acceptable and furthermore mandate that the justification is given within their DAS.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. RISE-2019758. This work began as a class project for the course "Developing Ethical and Responsible AI for the Earth Sciences" taught by A.M. at the University of Oklahoma, thus support for the project also comes from the University of Oklahoma.

Data availability statement. The information collected on data availability statements and the code utilized for Figs. 1-2 can be found at: `https://doi.org/10.5281/zenodo.13844985`. The word cloud for Fig. 3 is generated by `https://www.jasondavies.com/wordcloud/`.

References

Bertram, M. G., J. Sundin, D. G. Roche, A. Sánchez-Tójar, E. S. J. Thoré, and T. Brodin, 2023: Open science. *Current Biology*, **33 (15)**, R792–R797, https://doi.org/10.1016/j.cub.2023.05.036.

Briney, K. A., 2024: Measuring data rot: An analysis of the continued availability of shared data from a Single University. *PLOS ONE*, **19 (6)**, e0304 781, https://doi.org/10.1371/journal.pone.0304781.

Campbell, A., and J. Mu, 2023: Navigating Trust in Academic Research: The Rise of Data Availability Statements – Part I. *Digital Science*.

Edwards, P. N., M. S. Mayernik, A. L. Batcheller, G. C. Bowker, and C. L. Borgman, 2011: Science friction: data, metadata, and collaboration. *Social Studies of Science*, **41 (5)**, 667–690, https://doi.org/10.1177/0306312711413314.

Fecher, B., and S. Friesike, 2014: Open Science: One Term, Five Schools of Thought. *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*, S. Bartling, and S. Friesike, Eds., Springer International Publishing, Cham, 17–47, https://doi.org/10.1007/978-3-319-00026-8_2, URL https://doi.org/10.1007/978-3-319-00026-8_2.

Goodman, A., and Coauthors, 2014: Ten Simple Rules for the Care and Feeding of Scientific Data. *PLOS Computational Biology*, **10 (4)**, e1003 542, https://doi.org/10.1371/journal.pcbi.1003542.

Graf, C., D. Flanagan, L. Wylie, and D. Silver, 2020: The Open Data Challenge: An Analysis of 124,000 Data Availability Statements and an Ironic Lesson about Data Management Plans. *Data Intelligence*, **2 (4)**, 554–568, https://doi.org/10.1162/dint_a_00061.

Grant, R., and I. Hrynaszkiewicz, 2018: The Impact on Authors and Editors of Introducing Data Availability Statements at Nature Journals. *International Journal of Digital Curation*, **13 (1)**, 195–203, https://doi.org/10.2218/ijdc.v13i1.614.

Haupt, S. E., and Coauthors, 2022: The History and Practice of AI in the Environmental Sciences. *Bulletin of the American Meteorological Society*, **103 (5)**, E1351–E1370, https://doi.org/10.1175/BAMS-D-20-0234.1.

Liesenfeld, A., and M. Dingemanse, 2024: Rethinking open source generative AI: open-washing and the EU AI Act. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, 1774–1787, FAccT '24, https://doi.org/10.1145/3630106.3659005, URL https://doi.org/10.1145/3630106.3659005.

Maslej, N., and Coauthors, 2024: The AI Index 2024 Annual Report. Tech. rep., AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, 502 pp. URL https://aiindex.stanford.edu/report/.

McGovern, A., A. Bostrom, M. McGraw, R. J. Chase, D. J. Gagne, I. Ebert-Uphoff, K. D. Musgrave, and A. Schumacher, 2024: Identifying and Categorizing Bias in AI/ML for Earth Sciences. *Bulletin of the American Meteorological Society*, **105 (3)**, E567–E583, https://doi.org/10.1175/BAMS-D-23-0196.1.

McGuinness, L. A., and A. L. Sheppard, 2021: A descriptive analysis of the data availability statements accompanying medRxiv preprints and a comparison with their published counterparts. *PLoS ONE*, **16 (5)**, e0250 887, https://doi.org/10.1371/journal.pone.0250887.

Pineau, J., P. Vincent-Lamarre, K. Sinha, V. Lariviere, A. Beygelzimer, F. d'Alche Buc, E. Fox, and H. Larochelle, 2021: Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, **22 (164)**, 1–20.

Schuster, D., M. Mayernik, and G. Mullendore, 2023: Products developed through the ”What About Model Data?, Determining Best Practices for Preservation and Replicability, EarthCube Research Coordination Network” project. UCAR/NCAR - GDEX, URL https://gdex.ucar.edu/dataset/id/6962fde0-9f65-4530-9320-76c42866c821.html, https://doi.org/10.5065/G936-Q118.

Stall, S., and Coauthors, 2023: Ethical and Responsible Use of AI/ML in the Earth, Space, and Environmental Sciences. URL https://essopenarchive.org/users/536571/articles/635008-ethical-and-responsible-use-of-ai-ml-in-the-earth-space-and-environmental-sciences, https://doi.org/https://doi.org/10.22541/essoar.168132856.66485758/v1.

Tedersoo, L., and Coauthors, 2021: Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, **8 (1)**, 192, https://doi.org/10.1038/s41597-021-00981-0.

UNESCO, 2021: UNESCO Recommendation on Open Science. Tech. rep., UNESCO. https://doi.org/10.54677/MNMH8546, URL https://unesdoc.unesco.org/ark:/48223/pf0000379949.

Wilkinson, M. D., and Coauthors, 2016: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3 (1)**, 160 018, https://doi.org/10.1038/sdata.2016.18.

Wirz, C. D., and Coauthors, 2024: Increasing the Reproducibility and Replicability of Supervised AI/ML in the Earth Systems Science by Leveraging Social Science Methods. *Earth and Space Science*, **11 (7)**, e2023EA003 364, https://doi.org/10.1029/2023EA003364.