Open Science Expectations for Simulation-Based Research

Gretchen L. Mullendore*†, Matthew S. Mayernik† and Douglas C. Schuster†

National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), Boulder, CO, United States

There is strong agreement across the sciences that replicable workflows are needed for computational modeling. Open and replicable workflows not only strengthen public confidence in the sciences, but also result in more efficient community science. However, the massive size and complexity of geoscience simulation outputs, as well as the large cost to produce and preserve these outputs, present problems related to data storage, preservation, duplication, and replication. The simulation workflows themselves present additional challenges related to usability, understandability, documentation, and citation. These challenges make it difficult for researchers to meet the bewildering variety of data management requirements and recommendations across research funders and scientific journals. This paper introduces initial outcomes and emerging themes from the EarthCube Research Coordination Network project titled "What About Model Data? - Best Practices for Preservation and Replicability," which is working to develop tools to assist researchers in determining what elements of geoscience modeling research should be preserved and shared to meet evolving community open science expectations.

Specifically, the paper offers approaches to address the following key questions:

- How should preservation of model software and outputs differ for projects that are oriented toward knowledge production vs. projects oriented toward data production?
- What components of dynamical geoscience modeling research should be preserved and shared?
- What curation support is needed to enable sharing and preservation for geoscience simulation models and their output?
- What cultural barriers impede geoscience modelers from making progress on these topics?

Keywords: data, preservation, replicability, model, simulation

INTRODUCTION

Dynamical models are central to the study of Earth and environmental systems as they are used to simulate specific localized phenomena, such as tornadoes and floods, as well as large-scale changes to climate and the environment. High-profile projects such as the Coupled Model Intercomparison Project (CMIP) have demonstrated the potential value of sharing simulation output data broadly within scientific communities (Eyring et al., 2016). However, more focus is needed on open science

challenges related to simulation output. Researchers face a bewildering variety of data management requirements and recommendations across research funders and scientific journals, few of which have specific and useful guidance for how to deal with simulation output data.

Simulation-based research presents a number of significant data-related problems. First, simulations can generate massive volumes of output. Increased computing power enables researchers to simulate weather, climate, oceans, watersheds, and many other phenomena at ever-increasing spatial and temporal resolutions. It is common for simulations to generate tens or hundreds of terabytes of output, and larger projects like the CMIPs generate petabytes of output.

Second, interdependencies between hardware and software can limit the portability of models, and make the long-term accessibility of their output problematic. Many current data management guidance documents provided by scientific journal publishers conflate scientific computational models with software, thereby not addressing whether/how to archive model outputs. Equating computational models with software does not add much clarity to these recommendations, as ensuring "openness" of software is itself a significant challenge (Easterbrook, 2014; Irving, 2016). Models in many cases involve interconnections between community models, open source software components, and custom code written to investigate particular scientific questions. Large-scale models often also borrow and extend specific components from other models (Masson and Knutti, 2011; Alexander and Easterbrook, 2015).

Third, the lack of standardization and documentation for models and their output makes it difficult to achieve the goals of open and FAIR data initiatives (Stall et al., 2018). While this problem is not unique to simulation-based research, it has stimulated a number of initiatives to develop more consistency in how variables are named within simulation models, how models themselves are documented, and in how model output data are structured and described (Guilyardi et al., 2013; Heydebreck et al., 2020; Eaton et al., 2021).

The result is that the long-term value of simulation outputs is harder to assess than of observational data, and requires focused effort if the value is to be achieved. Key questions that challenge researchers who use such models are "what data to save" and "for how long?" Guidance on these questions is particularly vague and inconsistent across funders and publishers. "Reproducibility" is likewise difficult to define and achieve for computational simulations. Many different approaches have been proposed for what is required to successfully reproduce prior research (Gundersen, 2021). Within climate science, for example, bitwise reproducibility of model runs has not been a primary focus due to the non-linear nature of the phenomena being simulated, as well as the differences in bitwise output that occur when transferring models to different computing hardware (Bush et al., 2020).

Following the terminology of the recent US National Academies of Sciences, Engineering, and Medicine (2019) report on "Reproducibility and Replicability in Science," the primary goal in Earth and environmental science research is replicability of findings related to the physical system being simulated, not bitwise computational reproducibility. In other words, the goal

is to have enough information about research workflows and selected derived data outputs to communicate the important configurational characteristics to allow a future researcher to build from the original study.

This paper builds on the initial findings of the EarthCube Research Coordination Network (RCN) project titled "What About Model Data? - Best Practices for Preservation and Replicability" (https://modeldatarcn.github.io/) to address the following key questions related to open science and simulation-based research:

- How should preservation of model software and outputs differ for projects that are oriented toward knowledge production vs. projects oriented toward data production?
- What elements of dynamical geoscience modeling research should be preserved and shared?
- What curation support is needed to enable sharing and preservation for geoscience simulation models and their output?
- What cultural barriers impede geoscience modelers from making progress on these topics?

The goal of this discussion is to highlight initial findings and selected themes that have emerged from the RCN project. The discussion is not intended to provide prescriptive guidelines for what and how long data should be preserved and shared from simulation based research to fulfill community open science expectations. Instead, we share here initial progress toward guidelines, and, importantly, the broader themes that we have identified as crucial to understand and address in order to reach community open science goals. We plan to share detailed guidance for specific datasets in a future article.

RESEARCH COORDINATION NETWORK PROJECT OVERVIEW

The ultimate goal of the RCN project is to provide guidance on what data and software elements of simulation based research, specifically from dynamical models, need to be preserved and shared to meet community open science expectations, including those of funders and publishers. To achieve this goal, two virtual workshops were held in 2020, and ongoing engagement with selected stakeholders has taken place through professional society based town halls and webinars. Workshop participants included representatives from a variety of communities, including atmospheric, hydrologic, and oceanic sciences, data managers, funders, and publishers.

Project deliverables developed through the workshops and follow-on discussions include: (1) a preliminary rubric that can be used to inform a researcher on what simulation output needs to be preserved and shared in a FAIR aligned community data repository to support replicability of research results, and allow others to easily build upon research findings, (2) draft rubric usage instructions, and (3) an initial set of reference use cases, which are intended to provide researchers with examples of what has been preserved and shared by other projects that attained similar rubric scores. The current version of all project

deliverables can be accessed at https://modeldatarcn.github.io/. After further workshops are held and additional community input is gathered in 2022, project stakeholders plan to refine the project outputs further for later publication.

INITIAL RCN PROJECT FINDINGS

Knowledge Production vs. Data Production

A primary determinant of the data archiving for modeling projects is whether they are oriented toward knowledge or data production. Most scientific research projects are undertaken with the main goal of knowledge production (e.g., running an experiment with the goal of publishing research findings). Other projects are designed and undertaken with the specific goal of data production, that is, they produce data with the intention that those data will be used by others to support knowledge production research. For example, regional and global oceanic and atmospheric reanalysis products produced by numerical weather prediction centers would fall into the category of data production. The importance of this distinction is that different kinds of work are involved in knowledge vs. data production (Baker and Mayernik, 2020, **Figure 1**).

In particular, data production cannot occur without well-planned and funded data curation support, whereas knowledge production-oriented projects can be quite successful at generating new findings with minimal data curation. In some cases, such as the CMIPs, projects are designed for both knowledge and data production.

It is difficult to achieve either knowledge or data production if a project does not have that orientation from the beginning. Projects with a knowledge production orientation may generate significant amounts of data, and may want other scientists to use their outputs. But if data production is not the explicit goal and orientation from the beginning of a project, it is difficult for data to be used by others without direct participation by the initial investigator(s). If preservation and broad sharing of most project-generated data is intended to take place, a data production-orientation is necessary. This must encompass data preparation and curation tasks, such as ensuring that data and metadata conform to standards, that files are structured in consistent formats, that data access and preservation are possible, that data biases and errors are documented, and that data can be accessed and cited via persistent identifiers (McGinnis and Mearns, 2021; Petrie et al., 2021).

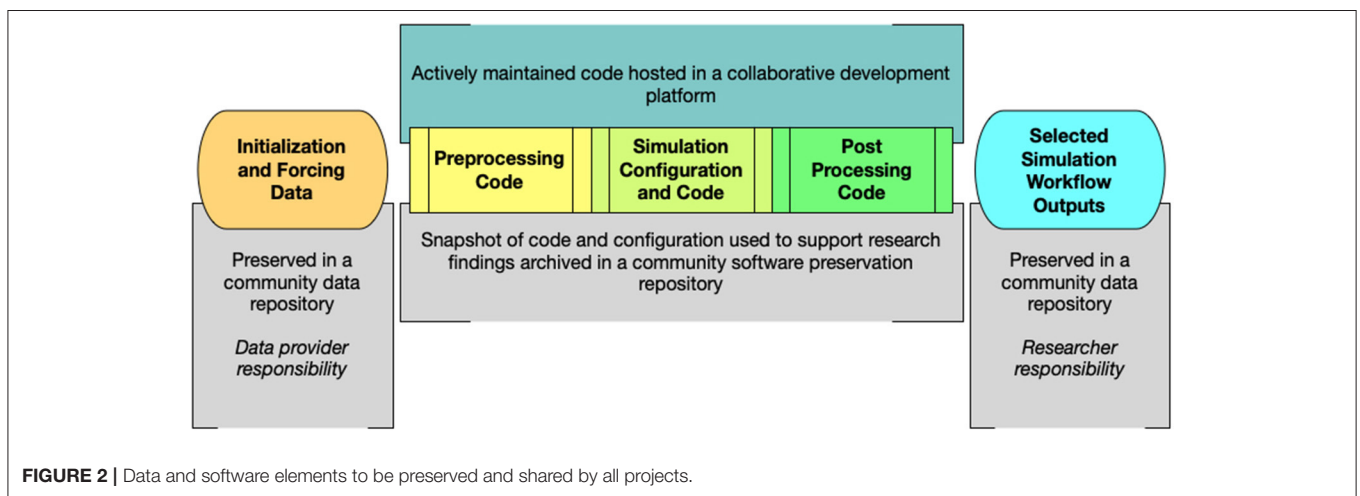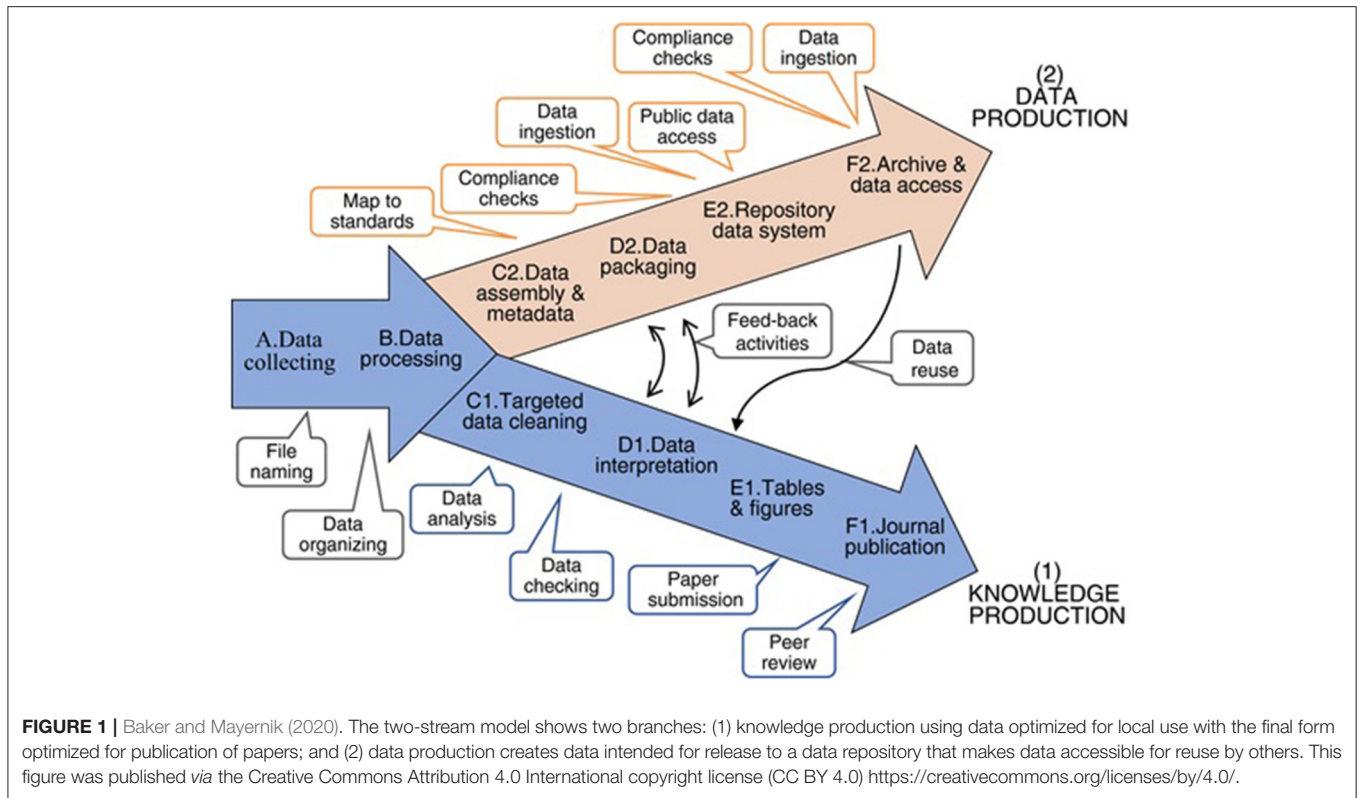## Determining What to Preserve and Share

While each project is unique, certain data and software elements should be preserved and shared for all projects to support research replicability and allow researchers to more easily build upon the work of others. Accordingly, workshop participants found that it would be best to preserve and share all elements of the simulation workflow, not just model source code (**Figure 2**). Simply sharing model code doesn't provide the level of understanding needed to easily build upon existing research. Also, if initialization and forcing data are provided by an outside provider, such as a national meteorological center, it should be the responsibility of that center to provide access to those data.

As discussed above, most scientific research projects are focused on knowledge production and as such should be saving little to no raw simulation data in repositories, instead focusing on smaller derived fields that help communicate to future researchers the environmental state or other information important for building similar studies in the future. Particularly for highly non-linear case studies, the goal is not exact reproducibility, but rather enough output to understand the environmental state that forced, and the impacts of, the features being investigated. There may be unique projects in which bitwise reproducibility is deemed necessary; in those cases, containerization can be useful (Hacker et al., 2016). However, to build upon prior research, most knowledge production research does not require bitwise reproducibility. Conversely, as described above, data production projects should have well-structured plans to preserve and share all model outputs needed for downstream users to successfully develop knowledge production research from those outputs.

Need for Curation Support

Development of research data and software that adheres to community best practice expectations for reuse requires specialized knowledge, and can be resource intensive. For example, data management includes a broad spectrum of activities in the data lifecycle, including proposal planning, data collection and organization, metadata development, repository selection, and governance (Wilkinson et al., 2016; Lee and Stvilia, 2017). Model code, output data, and any platforms being used to deliver code and/or software need to be documented clearly to provide guidance for potential users. Research software should be made available through collaborative development platforms such as GitHub (github.com) or Bitbucket (bitbucket.org), versioned, and licensed to describe terms of reuse and access (Lamprecht et al., 2020; American Meteorological Society, 2021). Both the data and snapshots of software versions that were used to support research outcomes should be archived in trusted data (e.g., https://repositoryfinder.datacite.org) and software (e.g., https://zenodo.org, https://figshare.com) repositories for long-term preservation and sharing, and assigned digital object identifiers to facilitate discovery and credit (Data Citation Synthesis Group, 2014; Katz et al., 2021).

The RCN project is working to develop strategies for deciding what needs to be preserved and shared, and communicate those practices clearly to researchers, repositories, and publishers. This should decrease the volume of simulation-related output that needs to be preserved, but conversely there is an expectation for researchers to share simulation configuration, model and processing codes that can reasonably be understood and reused by others with discipline specific knowledge. Researchers are currently spending a significant portion of their own time dealing with data curation; in some cases, over 50% of their funded time. Developing and stewarding software that adheres to community best practice expectations adds an additional burden on the researcher that may take up more of their funded time. Additionally, the availability of community data repositories in selected disciplines, such as the atmospheric sciences, is sparse, making it challenging for researchers to find an appropriate

FIGURE 1 | Baker and Mayernik (2020). The two-stream model shows two branches: (1) knowledge production using data optimized for local use with the final form optimized for publication of papers; and (2) data production creates data intended for release to a data repository that makes data accessible for reuse by others. This figure was published *via* the Creative Commons Attribution 4.0 International copyright license (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/.

FIGURE 2 | Data and software elements to be preserved and shared by all projects.

repository to deposit their data. A coordinated effort is needed to fund personnel to assist researchers in data and software curation, as well as investment in the needed repository preservation and stewardship services, to complement existing capabilities (Gibeaut, 2016; Mayernik et al., 2018). It is unreasonable to expect already overloaded researchers to become expert data managers and software developers, and find time to complete their research activities.

Cultural Barriers to Progress

As was discussed already, resources (time, money, personnel) remain a significant barrier to implementation of data

management best practices that promote increased scientific replicability, reduced time-to-science, and broadened participation. But there is also resistance to change as these practices are often in opposition to the way much of the community has built a successful career. Career advancement for scientists in typical scientific career pathways at research centers and universities is based on long-used metrics of "scientific success." The primary traditional metrics are number of publications, citations, and amount of proposals awarded. Often observational instrument researchers have built careers by leveraging use of their instrument in field campaigns to secure proposal dollars and subsequent publications. Some

model researchers and theorists argue that limiting access to their software is thereby an equivalent path that they take for building their career. However, instrument researchers are only a small subset of observationalists, and many scientists have built a strong career without limiting access to their software or data.

That said, we do recognize existing challenges in sharing of data and software. One challenge is the lack of adoption in formally citing datasets and software in peer reviewed journals, and consideration of such impact measurements in evaluations for promotion. Initiatives to add data and software contributions to the evaluation process for promotion and awards exist at various institutions, but adoption of new practices tends to be arduous. To protect early career scientists and researchers at smaller institutions, we propose using the practices of data and software sharing embargoes and curation waivers. Embargoes on data sharing are often used in field campaigns to give graduate students a certain amount of time to work with the data before sharing more broadly, in recognition of their need to publish on this data as part of their career development (and possibly needing more time to do so than more experienced researchers). This practice should be continued for new data and software to protect early career researchers. Additionally, waivers are often used to reduce requirements (e.g., publication fees) for researchers without sufficient resources. Here, we can extend waivers to curation requirements for researchers at institutions lacking in data/software curation expertise. Embargoes and waivers should not be used as excuses, however, to fall back on "data available upon request" statements that are proven to be problematic (Tedersoo et al., 2021). Overall, for these kinds of considerations, we emphasize not disproportionately punishing researchers with fewer resources.

Other common concerns from scientists about more open access, and particularly to software, are misuse of the software and fear of sharing suboptimal code. Misuse of software is a real outcome, as any open source software may ultimately be misused by some. However, the benefit of a more inclusive user base far outweighs the dangers of misuse (American Meteorological Society, 2021). A significant challenge is often determining who is responsible, if anyone, for user support, as this is rarely documented or formalized within research teams. As for sharing suboptimal code, most researchers in the Earth sciences are not formally trained programmers and many feel that their code is clunky, sometimes embarrassingly so. However, while some documentation is needed, elegant code is not a requirement for success in the earth sciences. In general, the community is accepting of code as long as it gets the correct physical answer. An added benefit of sharing code is that later users may streamline and optimize it, benefiting everyone.

DISCUSSION

We must work as a community to overcome the barriers to open data and software because our current practices impede broadened participation in the Earth sciences. Scientific equity cannot be fully achieved when individual scientists act as gatekeepers for new models, data, and software. However, these new initiatives need to be supported financially, with expertise and infrastructure, and incentivized through modernized merit review criteria (Moher et al., 2018). As discussed above, researchers are already struggling with data curation and code documentation and sharing; the community needs help from researchers trained in these areas (possibly as staff support at shared repositories). Without financial support and infrastructure provided for the scientific community, researchers at smaller institutions will be the hardest hit by these changes, negating the very advances we are trying to achieve in broadening participation. We can mitigate to some degree with embargos and waivers, but in the long run, we need federal commitment to data and software curation services.

Funding agencies are already paying for data work, if indirectly, by adding open data requirements to research grants but not increasing the investment in data infrastructures and data curation expertise. The result has been that scientists and graduate students re-allocate grant funding intended for scientific research to complete data tasks. If open science expectations for simulation-based research are to be achieved, the investment in data work should be more direct and intentional. Investigators spending research grant dollars on minimal curation by untrained graduate students is inefficient and will not lead to the intended outcomes of high-quality data sets being deposited in well-curated data repositories.

Finally, we emphasize that it is important to consider more than just the extremes for many of the questions and topics discussed in this paper. From our project's discussions, it is clear that we must get past the poles of either all or no model output being preserved. The best outcome in most use cases discussed within our project has been somewhere in the middle, namely that some output be preserved, but not all. Likewise, software need not be all open or closed. Some software may be released openly even if other software components are withheld from public view due to security or proprietary concerns. Similarly, questions about curation work should not be limited to a scientist vs. curator debate. Ideally, curation tasks should involve partnerships between scientific and data experts to take advantage of their respective knowledge and skills.

The next steps for our project and for the community broadly will be to address other important questions that have come up in our project activities, but have not been discussed in detail. For example, how long should simulation output be preserved and shared? Needs for data longevity are almost impossible to assess up front, due to the unknown future value and user bases of archived data sets (Baker et al., 2016). Such assessments have to be done downstream. But what are the best measures of a data set's value over time? Ideally this would be based on robust metrics, but there is not yet community agreement on what metrics are most appropriate. The overall goal is to make sure that we are preserving materials that can enable follow-on research, whether that be data, software, or both. More discussion and use cases will be necessary going forward to address these difficult challenges.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

GM was the Principal Investigator of the NSF-funded RCN project. MM and DS are Co-Principal Investigators.

All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Alexander, K., and Easterbrook, S. M. (2015). The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations. *Geosci. Model Dev.* 8, 1221–1232. doi: 10.5194/gmd-8-1221-2015

American Meteorological Society (2021). *Software Preservation, Stewardship, and Reuse: A Professional Guidance Statement of the American Meteorological Society.* Available online at: https://www.ametsoc.org/index.cfm/ams/about-ams/ams-statements/statements-of-the-ams-in-force/software-preservation-stewardship-and-reuse/

Baker, K. S., Duerr, R. E., and Parsons, M. A. (2016). Scientific knowledge mobilization: co-evolution of data products and designated communities. *Int. J. Digital Curat.* 10, 110–135. doi: 10.2218/ijdc.v10i2.346

Baker, K. S., and Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere* 11:3191. doi: 10.1002/ecs2.3191

Bush, R., Dutton, A., Evans, M., Loft, R., and Schmidt, G. A. (2020). Perspectives on data reproducibility and replicability in paleoclimate and climate science. *Harvard Data Sci. Rev.* 2:4. doi: 10.1162/99608f92.00cd8f85

Data Citation Synthesis Group (2014). *Joint Declaration of Data Citation Principles.* San Diego CA: FORCE. doi: 10.25490/a97f-egyk

Easterbrook, S. M. (2014). Open code for open science? *Nat. Geosci.* 7, 779–781. doi: 10.1038/ngeo2283

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., et al. (2021). *NetCDF Climate and Forecast (CF) Metadata Conventions.* Available online at: https://cfconventions.org/cf-conventions/cf-conventions.html (accessed October 11, 2021).

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., et al. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* 9, 1937–1958. doi: 10.5194/gmd-9-1937-2016

Gibeaut, J. (2016). Enabling data sharing through the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC). *Oceanography* 29, 33–37. doi: 10.5670/oceanog.2016.59

Guilyardi, E., Balaji, V., Lawrence, B., Callaghan, S., Deluca, C., Denvil, S., et al. (2013). Documenting climate models and their simulations. *Bull. Am. Meteorol. Soc.* 94, 623–627. doi: 10.1175/BAMS-D-11-00035.1

Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philo. Transact. R. Soc. A.* 379:2197. doi: 10.1098/rsta.2020.0210

Hacker, J., Exby, J., Gill, D., Jimenez, I., Maltzahn, C., See, T., et al. (2016). A containerized mesoscale model and analysis toolkit to accelerate classroom learning, collaborative research, and uncertainty quantification. *Bull. Am. Meteorol. Soc.* 98, 1129–1138. doi: 10.1175/BAMS-D-15-00255.1

Heydebreck, D., Kaiser, A., Ganske, A., Kraft, A., Schluenzen, H., and Voss, V. (2020). *The ATMODAT Standard enhances FAIRness of Atmospheric Model Data.* Washington, DC: American Geophysical Union. doi: 10.1002/essoar.10504946.1

Irving, D. (2016). A minimum standard for publishing computational results in the weather and climate sciences. *Bull. Am. Meteorol. Soc.* 97, 1149–1158. doi: 10.1175/bams-d-15-00010.1

Katz, D. S., Chue Hong, N., Clark, T., Muench, A., Stall, S., Bouquin, D., et al. (2021). Recognizing the value of software: a software citation guide [version 2; peer review: 2 approved]. *F1000Research* 9:1257. doi: 10.12688/f1000research.26932.2

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., et al. (2020). Towards FAIR principles for research software. *Data Sci.* 3, 37–59. doi: 10.3233/DS-190026

Lee, D. J., and Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE* 12:e0173987. doi: 10.1371/journal.pone.0173987

Masson, D., and Knutti, R. (2011). Climate model genealogy. *Geophys. Res. Lett.* 38:46864. doi: 10.1029/2011gl046864

Mayernik, M., Schuster, D., Hou, S., and Stossmeister, G. J. (2018). *Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop Report.* Boulder, CO: National Center for Atmospheric Research. doi: 10.5065/D6NC601B

McGinnis, S., and Mearns, L. (2021). Building a climate service for North America based on the NA-CORDEX data archive. *Climate Serv.* 22:100233. doi: 10.1016/j.cliser.2021.100233

Moher, D., Naudat, F., Cristea, I., Miedema, F., Ioannidis, J., and Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLoS Biol.* 16:e2004089. doi: 10.1371/journal.pbio.2004089

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science.* Washington, DC: The National Academies Press. doi: 10.17226/25303

Petrie, R., Denvil, S., Ames, S., Levavasseur, G., Fiore, S., Allen, C., et al. (2021). Coordinating an operational data distribution network for CMIP6 data. *Geosci. Model Dev.* 14, 629–644. doi: 10.5194/gmd-14-629-2021

Stall, S., Yarmey, L. R., Boehm, R., Cousijn, H., Cruse, P., Cutcher-Gershenfeld, J., et al. (2018). Advancing FAIR data in Earth, space, and environmental science. *Eos* 99:9301. doi: 10.1029/2018EO109301

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., et al. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* 8:192. doi: 10.1038/s41597-021-00981-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, I., J., Appleton, G., Axton, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18