EDITORIAL

Data Availability Principles and Practice

KEYWORD: Editorial

The American Meteorological Society (AMS) is moving toward strongly encouraging authors to make all data available, consistent with the "FAIR" principle: "findable, accessible, interoperable, and reusable." While we at the *Journal of Physical Oceanography* (*JPO*) firmly support this move, we also recognize that sharing all of the data is in some cases not practical or even useful. Your feedback now could help to prevent some of the less desirable possible side effects of this policy. We hope that this editorial will help to prod the discussion of exactly what data should be shared, and in what form (format, metadata, level of quality control, etc.).

The motivation is that science requires evidence. Making data available allows other scientists to confirm results, uncover errors, or find new insights. Moreover, gathering good data is expensive and time consuming. Since the same data can often be used for a range of purposes, making data available can be an efficient use of limited research resources. Doing so can also improve accountability when it comes to research findings. It is hoped that this would also mean that the originators of the data get full credit for its reuse [note that the National Science Foundation (NSF) now includes data archiving as a "product" along with the paper that is based on those data, so researchers *will* get credit there]. AMS recently updated its data policy guidelines (see https://www.ametsoc.org/PubsDataPolicy) to require, among other things, that papers in its journals include a Data Availability Statement. Data do not need to be completely "open" (although we encourage authors to make them as open as possible). Authors simply need to explain how to go about finding and using the data, or why, in some circumstances, the data cannot be made available.

We recognize that a "one size fits all" requirement on data availability could act to stifle some innovative work, particularly those involving very large datasets or novel instrumentation with data that cannot be constructively shared without a full course on the strengths and weaknesses of the approach. There may also be restrictions that are due to import/export regulations or other circumstances that prohibit free and full sharing of a dataset. To this point, the AMS guidelines do include some guidance. We present three pertinent examples:

- Example 1: Because of their proprietary nature ⟨or ethical concerns⟩, supporting data cannot be made openly available. Further information about the data and conditions for access are available at the [*repository name*] at [*insert DOI/URL here*].
- Example 2: Because of confidentiality agreements, supporting data can only be made available to bona fide researchers subject to a nondisclosure agreement. Details of the data and how to request access are available from [*data manager contact information*] at [*institution where data reside*].
- Example 3: Because of privacy and ethical concerns, neither the data nor the source of the data can be made available.

In addition to ethical and/or legal requirements, data availability is also subject to practical limitations. Data archiving resources may be insufficient to support those producing huge datasets. The online AMS guidance gives two more relevant examples:

- Example 4: The dataset on which this paper is based is too large to be retained or publicly archived with available resources. Documentation and methods used to support this study are available from [*data manager contact information*] at [*institution*].
- Example 5: The authors were unable to find a valid data repository for the data used in this study. These data are available from [*data manager contact information*] at [*host institution*].

This is certainly an issue for modelers, for whom the results of model runs can amount to many terabytes or more of output and could include many separate runs of the model (but note that model output can be replicated exactly by an identical model run). For example, model process studies often involve thousands of runs to explore the parameter space and to establish some uncertainty bounds.

Denotes content that is immediately available upon publication as open access.

DOI: 10.1175/JPO-D-20-0266.1

In such cases, it would likely be more useful (and practical) to point to the exact model and version that was used, along with a list of the exact setup files used.

This ''sheer volume'' problem also arises with many modern observational techniques, including high-speed video and 3D acoustic/optical approaches (e.g., scanning lidar, multibeam sonars, and satellite data; it is impractical to publicly archive multiple terabytes of data for each paper). Reasonable accommodation can be made in these cases—for example, by providing contact information for the person(s) taking care of the data.

The Data Availability Statement should make clear what has been archived and what steps have been taken to provide information about the data that could not be kept. A justification could describe the degree to which the documentation and methods provided should allow evaluation and replication of the study.

In this context, recent data openness and access initiatives within our community deserve appreciation. The Pangeo effort (https://pangeo.io/about.html) is making ocean and climate model data accessible using extremely efficient Internet access protocols. This effort, supported by the NSF, NASA, and the Sloan Foundation and initiated by Professor Ryan Abernathey at the Lamont–Doherty Earth Observatory, and others like it are developing the software infrastructure and providing pedagogical examples that make these data accessible to a much wider audience, including undergraduates and even high school students with programming experience. Models and modern observations produce vast quantities of data, and to date very few eyes have been laid upon them. Such projects will surely lead to an increase in the pace and quality of discoveries, but they cannot be established without the significant support of funding agencies, modeling centers, and universities. So, it is with care—based on all of the exceptions above to data sharing but also with the optimism that everyone will one day be able to set up data sharing easily and focus on the science of collecting data—that *JPO* is transitioning to an open data policy and acclaiming the present successes.

Thoughtful data availability requirements benefit both the scientific community and society. Consistent policies and practices can help to reduce misunderstanding and divergent interpretations. As editors, we do not wish or intend that the data availability requirement become a barrier to publication, whether because of the sensitivity of the data or because of limited resources. At the same time, the exceptions to making data available should not be used by researchers as a way to evade their responsibilities. We welcome authors and readers of *JPO* to look at the AMS data policy as posted and to contact us with any questions or concerns. To repeat, your feedback now could help to prevent some of the less desirable possible side effects of this policy, before it is fully implemented. For this reason, AMS is implementing this policy in steps, with increasing strictness, over the next year or so.

Acknowledgments. Much of this editorial is based on, and even paraphrased from, a similar editorial for *Weather, Climate, and Society* (Huntington et al. 2020).

Jerome A. Smith
Editor in Chief

Paola Cessi, *Ilker Fer*, *Gregory Foltz*, *Baylor Fox-Kemper*, *Karen Heywood*, *Nicole Jones*, *Jody Klymak, and Joseph LaCasce*
Editors

REFERENCE

Huntington, H. P., E. Archer, W. S. Ashley, S. L. Cutter, M. A. Goldstein, C. Roncoli, and T. L. Spero, 2020: Data availability principles and practice. *Wea. Climate Soc.*, **12**, 647–649, https://doi.org/10.1175/WCAS-D-20-0089.1.