PEER REVIEW OF DATASETS
When, Why, and How

by Matthew S. Mayernik, Sarah Callaghan, Roland Leigh, Jonathan Tedds, and Steven Worley

Devising methods for peer review of datasets can increase the trustworthiness and value of individual datasets and strengthen research findings.

Peer review holds a central place within the scientific communication system. Almost all forms of scientific work are subject to peer review, including journal articles, research grant proposals, and, in many research domains, conference papers and abstracts. Challenges to the peer review process are often seen as challenges to the integrity of science itself. In a vivid recent example, U.S. Representative Lamar Smith initiated congressional actions in the spring of 2013 that questioned the U.S. National Science Foundation’s (NSF) peer review process. In response, over 100 research and professional organizations signed a position statement defending NSF and the role of peer review in proposal assessments (www.aibs.org/position-statements/20130520_nsf_peer_review.html). While scholars of scientific communication have identified many potential biases that might affect the quality and neutrality of peer reviews [see Weller (2001) and Lee et al. (2013) for reviews of this literature], peer review is still recognized within most research institutions as the best method for evaluating the merits of scientific work.

A growing challenge to peer review is the increasing importance of digital datasets and computational research methods within scientific research. The increased emphases by funding agencies and research organizations are pushing scientific communities toward new approaches for data sharing, management, and preservation (Overpeck et al. 2011). Methods for assessing and ensuring data quality are also taking on new importance to engender trust in data and to enable secondary data use. Considerable challenges remain in incorporating data and methods into the peer review process. What Borgman noted in 2007 is still true today: “[T]he peer review of publications has few analogs for data. Questions loom about how data should be legitimized in the value chain [of scholarship] and at what point in their lifecycle that review should occur” (Borgman 2007, p. 131). LeMone and Jorgensen (2010), in their discussion of peer review within American Meteorological Society (AMS) journals, describe how datasets are often too big, too difficult to understand, poorly documented, or simply not accessible to would-be peer reviewers. Part of the challenge is that the notion of “data peer review” is itself vague and subject to different interpretations. The goal of this paper is to break down the idea of data peer review and illustrate how data peer review might be conceptualized and realized in different situations.

AFFILIATIONS: Mayernik and Worley—National Center for Atmospheric Research,* University Corporation for Atmospheric Research, Boulder, Colorado; Callaghan—British Atmospheric Data Centre, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Oxford, Didcot, United Kingdom; Leigh and Tedds—University of Leicester, Leicester, United Kingdom

CORRESPONDING AUTHOR: Matthew Mayernik, National Center for Atmospheric Research, University Corporation for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000
E-mail: mayernik@ucar.edu

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

The abstract for this article can be found in this issue, following the table of contents.
DOI: 10.1175/BAMS-D-13-00083.1
In final form 3 April 2014
©2015 American Meteorological Society
PREPARDE. This paper draws from the Peer Review for Publication and Accreditation of Research Data in the Earth Sciences (PREPARDE) project. PREPARDE was funded by Jisc (www.jisc.ac.uk/), a nonprofit organization in the United Kingdom that studies and promotes the use of digital technologies in education and research. PREPARDE investigated the processes and procedures required to publish scientific datasets, ranging from ingestion into a data repository through formal publication in a data journal, producing guidelines applicable to a wide range of scientific disciplines and data publication types. These guidelines are also informing the procedures and policies for authors, reviewers, and editors of articles being published by the Geoscience Data Journal, discussed more below (Callaghan et al. 2013).

PREPARDE project partners included organizations from both inside and outside the geosciences: from the United Kingdom, the University of Leicester, the British Atmospheric Data Centre, Wiley-Blackwell, the Digital Curation Centre, and Faculty of 1000, and, from the United States, the California Digital Library and the National Center for Atmospheric Research (NCAR). PREPARDE hosted town hall and splinter meetings at the 2012 annual meeting of the American Geophysical Union (AGU) and the 2013 annual meeting of the European Geophysical Union (EGU), as well as stand-alone workshops. The PREPARDE project also created a data publication mailing list (https://www.jiscmail.ac.uk/DATA-PUBLICATION) in February of 2013, which has since gained over 350 subscribers, and has been an active site for discussion.

BACKGROUND: DATA PUBLICATION AND CITATION. Interest in data peer review is tied to the development of new forms of data publication and citation. Data publication and citation initiatives are ongoing in multiple sectors, including within the atmospheric sciences. The AMS, for example, recently adopted a policy statement on “Full and Open Access to Data” that included a recommendation to develop a process for publishing and citing data referenced in AMS journals and publications (AMS 2013). The AMS Board on Data Stewardship,1 which led the writing of that statement, is actively investigating recommendations for the society to consider. Our approach follows AMS’s statement in discussing data publication as aligned with the principles and ethics of open science. Increased openness and availability of data benefits the broader scientific enterprise and increases the transparency and trustworthiness of scientific results.

Data “publication” can mean multiple things. Lawrence et al. (2011) distinguish between (capital “P”) data Publication and (small “p”) data publication. In their view, data Publication is the practice of making data “as permanently available as possible on the Internet” (p. 7), as well as putting data through processes that add value to the user, such as metadata creation and peer review. Data publication (small “p”), on the other hand, is simply the act of putting data on a website without a defined long-term commitment to digital archiving. Data Publication promotes the archiving of datasets as long-term resources that are stable, complete, permanent, and of good quality (Callaghan et al. 2012). If these criteria are met, datasets can be promoted as citable scholarly resources (Parsons et al. 2010; Mayernik 2013). As Parsons and Fox (2013) note, however, datasets challenge the “publication” concept. Many datasets are highly dynamic, regularly growing, or are duplicated, split, combined, corrected, reprocessed, or otherwise altered during use. The boundaries around a single dataset are variable across projects and organizations, leading to inevitable questions about the granularity of which data publications or citations should be designated. The inherent flexibility of the Internet, where resources can be posted, changed, and removed easily, ensures that these problems will be ongoing. Current recommendations on data publication and citation recommend taking combinations of technical and policy approaches, as neither work in isolation (see Socha et al. 2013; FORCE11 Data Citation Synthesis Group 2013).

Questions about data peer review naturally emerge when working toward data publication and citation. How do you know that a data publication is a valuable contribution? Does a data citation count toward hiring, promotion, and tenure decisions (which are traditionally based on assessing peer reviewed research)? The rest of this paper focuses on this data peer review issue.

DATA PEER REVIEW. Data peer review is not yet a well-defined process. Peer review can increase trust in scientific data and results and enable datasets to be evaluated and certified for quality. Data assessment processes and software, however, are often very specific to data types, experimental designs, and systems. In addition, scientists, data managers, and software engineers all have different expertise applicable to data peer review. Few reviewers would be qualified to review all aspects of a dataset. Many

---

1 Mayernik is currently a member of the AMS Board on Data Stewardship and Worley is a past Board member and coauthor of the AMS Full and Open Access to Data policy statement.
journals have policies stipulating that data should be available upon request, but in practice this can be difficult to enforce. Reviewers, while being as thorough as possible, must trust that authors are making good-faith attempts to be honest and complete in their descriptions of their data, data quality, and work processes (LeMone and Jorgensen 2010).

An additional impediment to data peer review is the fear from some would-be reviewers that data peer review would be extremely time intensive (Pampel et al. 2012). Peer review pipelines are already facing an exploding number of journal article submissions and grant applications (Miller and Couzin 2007). A recent survey by Golden and Schultz (2012) of reviewers for the publication *Monthly Weather Review* found that reviewers already review an average of eight journal articles per year, spending on average 9.6 h per review. No comparable study has been done specifically looking at the time and effort required to perform data peer review, although anecdotal evidence suggests that in some cases the this effort might be considerably less than 9.6 h (Callaghan 2015).

To address the scalability of data peer review as data volumes continue to increase, some academic communities are investigating how data peer review might be best conceptualized as a post publication process driven by data users (Kriegeskorte et al. 2012). Problems with data are often discovered after they are available for external use, regardless of the level of quality control in place up front (Hunter 2012). In the same sense that research publications receive post publication review via subsequent papers that attempt to replicate or verify particular findings, the true value of a dataset is often not established until post publication, after wide distribution occurs and a sufficient time period passes. Generally, over time the applicability of the dataset becomes more refined, as it is recognized to be appropriate for particular types of study but misleading if used in other types of study (see, e.g., http://climatedataguide.org; Schneider et al. 2013). What is abundantly clear is that with rapidly expanding data collections in all sectors of the sciences, adding data to the peer review system must be done deliberately and with forethought. The following section provides some guidelines informed by the PREPARDE project.

**DATA PEER REVIEW PRACTICES AND GUIDELINES.** The meaning of “data peer review” and the processes used (if at all) will vary by the kind of publication or resource being reviewed. The following sections outline considerations for data peer review in four scenarios: 1) data analyzed in traditional scientific articles, 2) data articles published in traditional scientific journals, 3) data submitted to open-access data repositories, and 4) datasets published via articles in data journals. These scenarios draw from an analysis by Lawrence et al. (2011) and represent common ways that datasets are presented, published, and archived in the geosciences. They constitute an ecosystem of venues for enabling data to be more widely discovered and used. No single category solves all of the problems related to digital data access and preservation. Instead, they should be thought of as complementary approaches, with data creators being able to leverage different options in different situations.

**Data analyzed in traditional scientific articles.** Most articles published in geoscience journals present analyses based on empirical data, often presented as charts, tables, and figures. For most journals, the review process only examines the data as they are presented in the articles, focusing on how such data influence the conclusions. For example, the reviewer guidelines for AMS journals, including the *Bulletin of the AMS (BAMS)*, do not discuss review of the underlying data (AMS 2014a,b). The AMS is not unique in this regard. Similarly, the AGU points reviewers to a recently published article in their publication *EOS, Transactions of the American Geophysical Union*, titled “A quick guide to writing a solid peer review” (Nicholas and Gordon 2011), which mentions data only in passing, simply noting that the data presented should fit within the logical flow of the paper and support the conclusions therein.

How should data peer review be approached for traditional scientific articles? Reviewing every dataset that underlies scientific articles is not a viable solution, and the merit and conclusions for many articles can be assessed without requiring a data peer review. With this in mind, journals need to aim for a solution that balances community accepted standards for validating scientific methods and authors’ and reviewers’ workloads. The first step is to improve author and reviewer guidelines so they reflect the science domain expectations for data transparency and accessibility. Through these guidelines, authors should be able to anticipate their community’s expectations. If the data are foundational to the merits of the published findings, it should be expected that the reviewers, editors, and fellow scientists will request access to the data. This access serves as an ongoing data peer review and supports additional scientific findings.

**Data articles published in traditional scientific journals.** In geoscience journals, it is common for articles to be published that announce the development of a new dataset or the release of a new version of an existing dataset. In fact, a search in the Web of Science citation index
in June of 2013 showed that 11 of the 20 highest cited articles ever published in BAMS can be categorized as data papers, in that their main focus was presenting a new or updated dataset to the BAMS community (see Table 1 for the list of papers). Usually these data papers provide a blend of descriptive characteristics about the data, scientific validation of the data quality, and comparison to other similar datasets. The peer review of such data papers follows the guidelines provided by the journal for all other papers, which, as discussed above, often do not have specific data-review guidelines.

When thinking about peer review for data papers published in traditional journals, the key element is persistence and longevity of the resource. Journals, in partnership with libraries, have established processes for ensuring the persistence and longevity of their publications. The expectation is the same for data presented in data papers. Thus, it is critical that the data and all of the relevant documentation are placed in a secure data repository with access methods suitable for the target community. In addition, the link between the paper and the data should be robust, as the paper provides an excellent (though not complete) source of documentation for the data. The use of persistent web identifiers, such as digital object identifiers (DOIs; http://doi.org), to identify and locate both data and articles increases the likelihood that those links are still actionable in the future.

As an example, the Research Data Archive at NCAR (RDA; http://rda.ucar.edu) strives to build strong connections between data papers in traditional journals and archived datasets. Data papers establish the dataset’s scientific credibility, and DOI referencing for the dataset enables these two research assets to be tightly coupled [e.g., the publication Large and Yeager (2009) describes the dataset Yeager and Large (2008); see also the publication Wang and Zeng (2013) and the dataset Wang and Zeng (2014)]. This linking of scholarly publications and datasets is growing rapidly in popularity. As metadata exchanges between journal publishers and data archives improve, along with increased dataset citation within publications, bibliographies of articles associated with datasets can be systematically compiled. This is a form of community data peer review.

Data submitted to open-access data repositories. In the geosciences, data repositories range from national data centers to project archives maintained by individual research organizations. In general, the kinds of data review that data repositories perform can be considered to be technical review, as opposed to scientific peer review. It should be noted, however, that data repositories are highly varied in their scopes, missions, and target communities. Depending on their purpose, some repositories may perform very minimal review of the datasets they archive, instead relying on the data providers to perform any review prior to data deposit.

The following discussion considers repositories that do perform data review. Such repositories typically perform a multistep technical review, participate in data quality assurance (QA) and data quality control (QC), and collaborate with the data providers and the scientific community. The major QA technical steps performed often include the following:

• Validating the completeness of the digital assets (data files and documentation).
• Evaluating the integrity of the data. The NCAR Research Data Archive, for example, has software that does accounting on data file content, examining every file to assure no corruption has occurred during production and transfer. This also ensures that the files contain exactly the data that are expected.
• Assessing the integrity of the documentation. The documentation can be textual or internal to the data files in the form of descriptive attributes. Metadata must be confirmed to be correct via automated processing or by visual inspection.

The information collected during the QA processes is often used to build informative and varied access services—for example, leveraging dataset metadata to enable effective repository searching and browsing.

Other QC processes, when performed, add supplementary information to data collections. This might include inserting quality flags, checking data values for consistency and known relationships (such as range checks based on physical limits), and validating or converting measurement units. Many of these processes require domain-specific analysis software and expertise to interpret the results.

Dedicated U.S. national repositories, like the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center (NCDC), have controlled procedures to manage the data deposits and carry out technical data reviews (see www.ncdc.noaa.gov/customer-support/archiving-your-data-ncdc). This includes collecting authors’ names and institutions, descriptions of the measured values, their relationship to other datasets and other versions of the same dataset, temporal and spatial coverage, file formats and access location for a data sample, data volumes, planned data transfer mechanisms, assessments of user community size, and other metadata. In general, the National Aeronautics and Space
Administration (NASA) Distributed Active Archive Centers (DAACs) have similar procedures whereby sufficient information is collected and a substantial technical data review is performed before the data are accepted. As another example, the Coupled Model Intercomparison Project phase 5 (CMIP5) has a multi-step quality-control process in place to ensure that the data contained within its archive conform to metadata requirements, are consistent in structure, and will be accessible and usable over time (see https://redmine.dkrz.de/collaboration/projects/cmip5-qc/wiki).

It is also common practice for repositories to provide mechanisms for feedback loops between users, the archive, and data providers, because QA and QC work are often stimulated by user questions. On these occasions, repository personnel must validate the suspicious user finding and collaborate with the data provider to understand the issue and determine a solution. At a minimum, the existence of suspicious data needs to be well documented, along with any dataset changes or new versions. Critically, however, repositories almost always maintain the original data to be able to produce the original resource by necessity or upon request.

In addition to providing data, repositories have considerable expertise that can be leveraged to develop data management plans. Evaluating and testing preliminary data output for compliance to standards and requirements allow for early discovery of problems and, ultimately, higher-quality data. These choices determine how well the data can be ingested, documented, preserved, and served to the broader scientific community. Preparatory work before data collection reduces subsequent QA and QC problems and simplifies the task of any data peer reviewer or data user down the line. Data providers and repository personnel should more aggressively perform preproject collaborations; this has a minor impact on resources, yet enables significant efficiencies and benefits downstream.

Datasets published via articles in data journals. Data journals, a fairly recent development in scientific publishing, primarily (or exclusively) publish data articles—that is, articles that describe research datasets and are cross-linked to datasets that have been deposited in approved data centers. A data paper describes a dataset, giving details of its collection, processing, software, and file formats, without the requirement of analyses or conclusions based on the data. It allows the reader to understand when, how, and why a dataset was collected, what the data products are, and how to access the data. Data papers are a step toward providing quality documentation for a dataset and begin the feedback loop between data creators and data users. Like any paper, data papers might contain errors, mistakes, or omissions and might need to be revised or rewritten based on the recommendations of peer reviewers.

Three data journals are discussed here as examples: Earth System Science Data, Geoscience Data Journal, and Scientific Data. Earth System Science Data (ESSD), published by Copernicus Publications, has published data articles since 2009 (www.earth-system-science-data.net). As of March 2014, over 80 articles have been published in ESSD since its inception. ESSD articles are first made open for public comments as “discussion papers” and are subsequently published in “final” form after completing the peer review and revision processes. The Geoscience Data Journal (GDJ), published by Wiley (www.geoscientificdata.com), began accepting submissions in late 2012 and, as of March 2014, has published eight articles. GDJ differs from ESSD in that the data articles in GDJ do not go through a discussion-paper phase before peer review or publication, and GDJ data papers may be published about datasets that are not openly available to everyone. The Nature Publishing Group is also developing a data journal, called Scientific Data, which was launched in May 2014 (www.nature.com/scientificdata). Unlike ESSD or GDJ, Scientific Data has a broad disciplinary focus, initially focused on the life, biomedical, and environmental sciences. Notably, Scientific Data is calling its publications “data descriptors,” not data papers or articles. The key difference between Scientific Data and ESSD or GDJ is that each data descriptor will include a structured metadata component in both human and machine-readable form, in addition to the narrative component of the publication.

These three data journals have considerable overlap in their guidelines for peer reviewers. These guidelines are summarized in Table 2. All three emphasize that reviewers assess the completeness of the dataset, the level of detail of the description, and the usefulness of the data. Usefulness is a difficult criterion to apply in practice, since predicting how data might be used in the future is very difficult for data creators or would-be reviewers. Thus, “usefulness” is typically discussed in relation to how the data paper can enable both intended and unintended uses of the data, as well as replication or reproduction of the data. Reviewers are also best positioned to provide feedback on whether data papers are written at the appropriate dataset granularity, as they understand community expectations for how data should be presented and are likely to be used. Data papers about a single weather station, for example, are likely to have little utility for the broader community, unless a particular station
was notable for historical or event-specific reasons. Reviewers, as members of the relevant scientific community, should conduct this assessment.

In addition, all three journals emphasize that the review assess the openness and accessibility of the data. Each journal partners with external data repositories who host the data described in the data papers; none host the data themselves. The journals provide lists of suggested or approved repositories, along with lists of repository requirements. The details of these requirements are generally consistent. Repositories must (i) assign persistent identifiers to the datasets, (ii) provide open public access to the data (including allowing reviewers prepublication access if required), and (iii) follow established methods and standards for ensuring long-term preservation and access to data. The partnerships with established and approved data repositories provide an extra level of quality assurance for the data and data paper.

**DISCUSSION—OVERARCHING CONSIDERATIONS.** All four of these publication types add to the body of knowledge that grows and surrounds research data. While they represent different venues for publishing data, they have some commonalities and differences with regard to data peer review and quality assurance of data.

**Commonalities between data publication types.** The first commonality is the need for data accessibility. Datasets cannot be reviewed if they are not available to the reviewers. Most data repositories provide open access to data submitted to them, and data journals address this issue by explicitly requiring that data are archived and available via a data center or repository. Few traditional journals in the geosciences, however, have instituted such data deposition requirements for a variety of reasons, which challenges organized data peer review. The second commonality is that authors are responsible for providing enough information for the dataset to be reviewed. Traditional journal articles provide an excellent source of data documentation and analysis, but space considerations limit the amount of detail that authors can provide. Data repositories rely on data providers to provide sufficient metadata for the dataset to be archived, preserved, and used properly. Data journals use the data paper as a primary source of data documentation and leverage the additional metadata provided by the repository hosting the data. The third commonality is that data peer reviewers need clear guidelines for how to perform a data review and for what characteristics of a dataset should be examined. Data peer review is a new-enough topic that few broad scale guidelines have been issued. The guidelines produced by data journals, shown in Table 2, are currently the most detailed.

**Differences between particular data publication types.** The most straightforward distinction in data-review processes is between the journal-based data publications and data repositories. Data review within a data repository primarily concentrates on technical aspects of the dataset in order to ensure that the dataset can be managed and curated properly. Successful completion of a data review within a data repository also provides an initial indication of whether the dataset is understandable by someone other than the dataset creator. However, the value of the dataset to the scientific community needs to be judged by that community.

The next notable difference is that, unlike data journals, most traditional journals currently do not ask peer reviewers to review in depth the data underlying the scientific findings. Data articles published in traditional journals normally do a careful scientific assessment of the data quality as compared to other similar datasets or known environmental conditions. This comparative and validation work is based on science and supports data quality, defines how the data should be used, and identifies uncertainties in the data products being presented. In announcing their dataset to the community, authors also typically include pointers to where the data can be acquired. Data articles in traditional journals are excellent knowledge anchor points for datasets. Reviewers are expected to assure the archived information meets the needs and expectations of the target community.

**Considerations of tools and processes for peer review.** A number of tools and processes might help in establishing data peer review practices. In general, the first step of the data review process is to validate the existence, access procedures, and completeness of the associated metadata. It is naive to assume that the initial dataset publication at a repository is always 100% perfect. Some data reviewers may prefer to download all or a representative portion of a dataset and use their own tools to examine the content and check the integrity. This is becoming less necessary, and in fact challenging for large (terabyte-sized) collections, because data repository portals are increasingly providing users in-line tools, such as quick-look viewers for the data, allowing plotting and overplotting of particular datasets. Standard sets of statistical tools to assess particular data types also enable reviewers to easily perform common analyses, such as trend analysis for time series and spatial variability analysis.
for geospatial datasets. Reviewers might also benefit from the ability to subset the data to explore particular data components.

To simplify the process of linking articles with underlying data, journals might partner with data repositories to enable researchers to submit data to a repository alongside their submission of an article to a journal. This approach has proven successful in the ecology and evolutionary biology community, where the Dryad data repository (http://datadryad.org/) partners with 50+ journals to provide an integrated pipeline for archiving data associated with journal articles.

Another approach might be for journals or data repositories to use a rating system to indicate that particular quality control or peer review processes have taken place. Costello et al. (2013) present an approach for such a rating system, with ratings ranging from one star, indicating that data have been submitted with basic descriptive metadata, to five stars, which might indicate that automated and human quality-control processes have been conducted, along with independent peer review, and an associated data paper has been published. This also supports ongoing efforts to link datasets and scholarly publications; as a dataset is cited more and more often, the community is validating its relevance and quality and thus the rating and value increases.

Recommendations for the AMS. Scientific professional societies have important leadership roles in advancing scientific practice and educating their communities. The AMS, for example, arguably crosses more institutional boundaries in the atmospheric and meteorological sciences than any other organization in the United States, spanning the commercial, academic, and government sectors. The AMS “Full and Open Access to Data” statement is an important leadership step in encouraging and enabling open access to data and increased scientific integrity. The following recommendations present a number of ways in which the AMS can take additional steps along this path and provide the maximum benefits for the scientific community, with the least increase in cost and effort required to publish a paper or conduct peer review for AMS journals.

1) Add data peer review recommendations to author guidelines. At minimum, authors should be referred to and understand the principles outlines in the AMS “Full and Open Access to Data” statement. Author guidelines should also discuss how datasets should be cited and linked in order to enable data access and transparency—for example, recommending that papers include a verifiable data citation that provides an access pathway to the data and metadata or a statement of the fact why the data cannot be made available (e.g., for ethical, legal, or privacy reasons).

2) Add data peer review recommendations to peer review guidelines. These guidelines can build on the efforts of data journals, shown in Table 2. Not all papers will need data peer review, but reviewers should have a clear set of steps to follow in determining whether data peer review is necessary and in how to conduct such a review.

3) Encourage data creators and data repository personnel to engage earlier in developing partnerships and sharing expertise. Data creators can, with a small amount of planning, prepare collections that use standard formats and have adequate metadata, smoothing the process for ingestion into repositories. AMS can facilitate these connections by providing a list of recommended data repositories, including both discipline-specific and discipline-agnostic data repositories. AGU has developed a data repository list that provides a good starting point for consideration (see http://publications.agu.org/files/2014/01/Data-Repositories.pdf).

4) AMS should formally endorse other efforts within the scientific community to encourage data citation, specifically the Earth Science Information Partners (ESIP) Federation’s guidelines (http://commons.esipfed.org/node/308) and the international principles on data citation (Socha et al. 2013; FORCE11 Data Citation Synthesis Group 2013).

Transition to tightly coupled data and scholarly work in AMS publications will take time. However, now is a reasonable time to set a schedule for an improved publication process supported by new AMS guideline documents, increased engagement with data repositories, and the development of educational materials and tutorial sessions at meetings.

Conclusions. Data peer review is not a monolithic concept. Different methods of presenting, publishing, and archiving data will require differing approaches to reviewing data. The four publication scenarios discussed in this paper illustrate how data peer review currently differs between traditional methods of presenting data in scientific articles, data

---

2 The AMS authors guideline was updated in early 2015 to include recommendations on data citation. Two authors, Mayernik and Worley, contributed to the development of these new guidelines.
papers written specifically for data journals, and data archived within open-access data repositories. Most journals do not provide reviewer recommendations or requirements related to the data, even though data papers are commonly published in geoscience journals and are often highly cited. Data repositories perform technical review of datasets as part of the process of data archiving, preservation, and service development. Data journals, which publish primarily data papers, have the most well-specified data peer-review processes and partner with data repositories for data access and archiving.

Looking ahead at the development of data peer review, three major issues need to be addressed. First, the accessibility of data underlying scientific articles is still highly variable. Data journals are able to develop thorough data peer review procedures because they partner with data repositories to ensure that data will be available to reviewers and users alike. Scientific journals, on the other hand, have highly variable policies on data archiving and have many competing interests to balance when creating such policies, only one of which is data accessibility. Nevertheless, there would be an overall benefit to scientific integrity and progress if journal articles were peer reviewed with criteria focused on accurate data citation and accessibility. Second, data peer review requires different expertise than peer review of traditional articles. Scientific expertise is only one consideration for data peer review. Knowledge of data structures and metadata standards need to be applied when evaluating datasets. Expertise in the data collection method and instrumentation is also relevant. In short, the pool of data peer reviewers should have a wider disciplinary distribution than the pool for typical scientific articles. The last major issue to be

<table>
<thead>
<tr>
<th>Table 1. Most cited BAMS articles. Data from Web of Science, gathered 11 Jun 2013.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>17</td>
</tr>
</tbody>
</table>
addressed is the pre- versus post publication review question. As data volumes continue to grow at exponential scales, prepublication peer review may need to be applied more selectively to data than to articles. Postpublication review in the form of comments, metrics of downstream data use, and data revision may prove to be more scalable in terms of people, time, and resources. Postpublication review might also leverage automated processes. For example, if data citations using DOIs became standard in all journal publications, indexing systems could collect lists of journal articles that cite a particular dataset. Links to those articles could also be automatically added to the dataset reference list maintained by the data repository. This would tie the understanding and impact gained from the dataset to the dataset itself. As the many data publication initiatives evolve, the most effective recommendations and practices will emerge.

Table 2. Data journal peer review guidelines. (Note: these are drawn from the associated websites. Some edits have been made for presentation.)

<table>
<thead>
<tr>
<th>Earth System Science Data</th>
<th>Geoscience Data Journal</th>
<th>Scientific Data</th>
</tr>
</thead>
</table>

I. Read the manuscript:
1. Are the data and methods presented new?
2. Is there any potential of the data being useful in the future?
3. Are methods and materials described in sufficient detail?
4. Are there any references/citations to other datasets or articles missing or inappropriate?

II. Check the data quality:
5. Is the dataset accessible via the given identifier?
6. Is the dataset complete?
7. Are error estimates and sources of errors given (and discussed in the article)?
8. Is the accuracy, calibration, processing etc. state of the art?
9. Are common standards used for comparison?

III. Consider article and dataset:
10. Are there any inconsistencies within these, implausible assertions or data or noticeable problems which would suggest the data are in error (or worse).
11. If possible, apply tests (e.g., statistics).
12. Unusual formats or other circumstances which impede such tests as are usual in your discipline may raise suspicion.

IV. Check the presentation quality:
13. Is the dataset usable in its current format and size?
14. Is the formal metadata appropriate?

Finally:
By reading the article and downloading the dataset would you be able to understand and (re-)use the dataset in the future?

I. Data description document
1. Is the method used to create the data of a high scientific standard?
2. Is enough information provided (in metadata also) to enable the data to be re-used or the experiment to be repeated?
3. Does the document provide a comprehensive description of all the data that is there?
4. Does the data make an important and unique contribution to the geosciences?
5. What range of applications to geosciences does it have?
6. Are all contributors and existing work acknowledged?
7. Does the Data Paper contain sufficient information about the dataset, e.g., dataset DOI, name of data center etc.?

II. Metadata
8. Does the metadata establish the ownership of the data fairly?
9. Is enough information provided (in data description document also) to enable the data to be re-used or the experiment to be repeated?
10. Are the data present as described, and accessible from a registered repository using the software provided?

III. The data themselves
11. Are the data easily readable, e.g. do they use standard or community formats?
12. Are the data of high quality e.g., are error limits and quality statements adequate to assess fitness for purpose, is spatial or temporal coverage good enough to make the data useable?
13. Are the data values physically possible and plausible?
14. Are there missing data that might compromise its usefulness?

I. Experimental Rigor and Technical Data Quality
1. Were the data produced in a rigorous and methodologically sound manner?
2. Was the technical quality of the data supported convincingly with technical validation experiments and statistical analyses of data quality or error, as needed?
3. Are the depth, coverage, size, and/or completeness of these data sufficient for the types of applications or research questions outlined by the authors?

II. Completeness of the Description
4. Are the methods and any data processing steps described in sufficient detail to allow others to reproduce these steps?
5. Did the authors provide all of the information needed for others to reuse this dataset, or integrate it with other data?
6. Is this Data Descriptor, in combination with any repository metadata, consistent with relevant minimum information or reporting standards?

III. Integrity of the Data Files and Repository Record
7. Have you confirmed that the data files deposited by the authors are complete and match the descriptions in the Data Descriptor?
8. Have these data files been deposited in the most appropriate available data repository?
ACKNOWLEDGMENTS. Many thanks to the PREPARDE partners and participants for their contributions to the project. We also thank Mary Marlino and the three peer reviewers for comments on previous drafts. The support of Jisc and the U.K. Natural Environment Research Council in funding and supporting the PREPARDE project is gratefully acknowledged.

REFERENCES


———, R. Lowry, and D. Walton, 2012: Data citation and publication by NERC’s Environmental Data Centres. Ariadne, 68. [Available online at www.ariadne.ac.uk/issue68/callaghan-et-al.]


FORCE11 Data Citation Synthesis Group, cited 2013: Joint declaration of data citation principles—Final. [Available online at www.force11.org/datacitation.]


Lawrence, B., C. Jones, B. Matthews, S. Pepler, and S. Callaghan, 2011: Citation and peer review of data: Moving towards formal data publication. Int. J. Digital Curation, 6, 4–37, doi:10.2218/ijdc.v6i2.205.


——, and ——, 2014: Global hourly 0.5-degree land surface air temperature datasets. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, doi:10.5065/D6PR7SZF.


