

### Public Domain Mark 1.0 Universal

This work was written as part of one of the author's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law.

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

### **Please provide feedback**

Please support the ScholarWorks@UMBC repository by emailing [scholarworks-group@umbc.edu](mailto:scholarworks-group@umbc.edu) and telling us what having access to this work means to you and why it's important to you. Thank you.

## Improving Reproducibility in Earth Science Research

*Earth scientists need software technology that better integrates legacy data with current and future processing capabilities so they can assess and reproduce their colleagues' results.*

By Zhong Liu, J. Wang, S. Pan, and David Meyer

30 October 2019



Software-based workflow management systems that incorporate standards from the Earth science community can facilitate the assessment of repeatability, replicability, and reproducibility of scientific claims and bridge current and future computing environments. Credit: geralt, Pixabay License

A cornerstone of solid science is the ability of scientists to assess the correctness of other researchers' results and conclusions critically and without restrictions [see [Plessner, 2018](#); [National Academies of Sciences, Engineering, and Medicine, 2019](#)]. Three common practices for such assessments, often called the 3Rs, range in difficulty from low to high [[Association for Computing Machinery, 2016](#)]:

- repeatability (same team, same experimental setup)
- replicability (different team, same experimental setup)
- reproducibility (different team, different experimental setup)

Much of Earth science today is computationally heavy, involving the use of specialized algorithms, software, and [computing environments](#). Reproducing such science requires that not only the software but also the associated data and information about the computing environment that generated the original results be available to other researchers. In reality, this is difficult for a number of reasons. Here we discuss these challenges and address how new software technology could better facilitate scientific reproducibility in Earth science.

## Mapping the Path from Data to Results

Scientific investigations usually follow a workflow—a sequence of steps through which data are processed and analyzed to give an end result, or product. Examples of Earth science workflows include relatively straightforward analyses of sea surface temperatures to sophisticated numerical modeling of weather and climate. Numerous computer programs and software packages have been developed to support scientific research workflows and their applications in Earth science. For example, complex computer software uses data from spaceborne sensors and other Earth observations to retrieve environmental parameters such as precipitation and hydrometeor (liquid or solid water particle) profiles.

“

***Access to input data is fundamentally important for reproducibility.***

Access to input data is fundamentally important for reproducibility. When a scientist or a journal reviewer tries to reproduce someone else's results, it can be challenging to locate the data used, especially if the research was not done recently.

Although publishers increasingly require authors to include source information for the data used in their research (e.g., web links of last access and digital object identifiers (DOIs)), problems remain. Because many Earth science disciplines [generate data sets](#) with different formats, data structures, and file stitching or aggregation methods, data sets quoted from web links or DOIs may not be immediately usable, instead requiring preprocessing that involves a lot of work and technical expertise. But reviewers are volunteers who seldom have much time to spend on data processing, especially if input data sets are large.

Reproducing research results also requires that any software used in generating the results is available to other scientists. This is not always the case, however. And even when it is available, missing information or incomplete descriptions can make the software hard to understand.

The workflow embedded in software must be well described for others to understand how it processes data. This description includes input and output data sets, workflow logic, algorithms used, the version of the software or library used, and more.

But researchers are under immense pressure to publish their work and do not always have time to devote to documenting or training outside researchers in their programs and processes. Furthermore, software packages are often written by students, postdocs, or interns, who are often not available to provide continuing support for their software after they complete their studies and move on to other institutions. So researchers seeking to reproduce results generated using custom-built software often find that support services are not available to answer questions.

Another major cause of missing software or incomplete documentation is that scientific publishers generally do not require that authors submit and publish custom software or code they used in their work. Therefore, scientists have very little incentive for doing the extra work to make the software publicly available. On the other hand, even if a scientist submits software to a publisher along with a research paper [e.g., [Science](#), 2019], in practical terms it is still very difficult for reviewers to retrieve input data, understand, and successfully run the software in their own computing environment.

## Supporting Legacy Data and Computing Environments

For older research papers, [data access](#) problems are even bigger. Who will ensure that the data and software used in these papers are still available for reproducibility after a number of years? Data DOIs are used in some research papers to reference the data involved, but eventually, it is the responsibility of data archive centers to uninterruptedly maintain data and services.

For example, there can be multiple versions of satellite-based data products because algorithms evolve and are improved over time. A common practice for data archive centers is to keep only the latest version of a data product, which can make accessing data sets in earlier versions very difficult, especially for products derived from raw measurement data.

Scientists seeking to assess or verify results in old publications face an even more difficult challenge. In theory, scientific data processing centers or principal investigators can reproduce these old-version data products using algorithms with raw measurement data. In reality, limited resources and limited demand for legacy data products

“

*Who will ensure that the data and software used in research papers are still available for reproducibility after a number of years?*

compared with demand for the newest version make this a difficult task. Thus, it is necessary to archive both raw and derived data sets in all versions because previously published research papers are linked to these data sets.

In addition to software and data, scientific reproducibility requires knowledge of and access to the required computing environment. This environment includes the appropriate computer operating system (in the version used to generate the original output), sufficient [data storage resources](#), adequate computing power, and the necessary software libraries (which can be outdated or missing entirely). Scientists often find that the computing environment they have cannot support software from the third party that generated the data product.

Compounding these problems are the security risks associated with running third-party software, even software provided by fellow scientists that can impose threats unknowingly. These risks can be dangerously high, making scientists reluctant to run third-party software.

## Updating Software Technology

“

***Earth scientists need a software-based workflow management system to liberate them from tedious computer hardware and software tasks and allow them to focus more on science issues.***

Can software technology help remove barriers to reproducibility in the Earth sciences, given the many complicated requirements for reproducibility? Earth scientists, who are often not computer experts, need a software-based workflow management system to keep all the related elements organized. Such a system can liberate them from tedious computer hardware and software tasks and allow them to [focus more on science issues](#) [[Claerbout and Karrenbach](#), 1992; [Donoho et al.](#), 2009; [Peng](#), 2011].

Workflow management software is important for efficiently and successfully implementing the 3Rs. A management system should be able to track the progress of each workflow automatically and record detailed information about the data, software, and computing environment. The system should also record and track all activities involved in each step of scientific processing, such as data inputs and outputs, data analysis, and visualization. Recorded provenance information can be attached with journal paper submissions so that peer reviewers or other colleagues can examine the details and run the workflow independently.

The workflow management system must be user friendly to minimize the learning curve and maximize its usefulness in 3R activities. The system must also enable scientists to conduct collaborative work more efficiently in different communities by making reproducibility not only possible but also simple.

Such a system benefits data archive centers as well because the centers can record and provide provenance information for their archived data sets. Scientists who download data from these data centers can then pass this information along



and add it to new workflows.

## Enabling the 3Rs

Several workflow management systems already exist [e.g., [Kepler](#), 2019] that allow scientists to add different analytical methods and to record the workflow provenance. However, these systems still have many limitations. We argue that it is critical to remove these obstacles and simplify the process for implementing the 3Rs.

Currently, for example, provenance information may not be interoperable from one workflow system to another, and there may not be sufficient provenance information available to perform the 3Rs [e.g., [World Wide Web Consortium](#), 2013]. Therefore, Earth science community stakeholders need to develop 3R standards. Furthermore, most existing systems require users to have expertise in computer science to add analytical algorithms into a workflow, but most Earth scientists do not currently have the requisite level of expertise.

We envision that future computing in Earth science will occur in an integrated environment, most likely based on cloud computing. In such an environment, scientists can run software and do data analysis “close to the data” using the same shared resources rather than downloading data sets to their own computing environments. In such an environment, standard provenance information will be automatically recorded for each run.

“

*To overcome these challenges, it is necessary to develop software management systems with community-based standards to bridge current and future computing environments.*

However, until this happens, we need to bridge current and future software practices. For example, to better document their research, scientists need workflow systems that can automatically generate provenance information based on community-defined standards. A scientist could then export the standardized and human-readable provenance information to their paper for journal submission. The system, meanwhile, could also assemble the software code, input data information, and other information into a virtual package like a [Docker](#) image so the package could be deployed seamlessly by other scientists.

There are many challenges in enabling repeatability, replicability, and reproducibility in Earth science. To overcome these challenges, it is necessary to develop software management systems with community-based standards to bridge current and future computing environments. New mandatory requirements from stakeholders will likely play an important role in accelerating the development of such systems and community-based standards. These systems, especially if they prove user-friendly, will help facilitate the 3Rs.

## References

Association for Computing Machinery (2016), Artifact review and badging, [www.acm.org/publications/policies/artifact-review-badging](http://www.acm.org/publications/policies/artifact-review-badging).

Claerbout, J. F., and M. Karrenbach (1992), Electronic documents give reproducible research a new meaning, *SEG Tech. Program Expanded Abstr.*, 11, 601–604, <https://doi.org/10.1190/1.1822162>.

Donoho, D. L., et al. (2009), Reproducible research in computational harmonic analysis, *Comput. Sci. Eng.*, 11(1), 8–18, <https://doi.org/10.1109/MCSE.2009.15>.

Kepler (2019), The Kepler Project, [kepler-project.org](http://kepler-project.org).

National Academies of Sciences, Engineering, and Medicine (2019), *Reproducibility and Replicability in Science*, Natl. Acad. Press, Washington, D.C., <https://doi.org/10.17226/25303>.

Peng, R. D. (2011), Reproducible research in computational science, *Science*, 334, 1,226–1,227, <https://doi.org/10.1126/science.1213847>.

Plessner, H. E. (2018), Reproducibility vs. replicability: A brief history of a confused terminology, *Front. Neuroinf.*, 11, 76, <https://doi.org/10.3389/fninf.2017.00076>.

Science (2019), *Science* journals: Editorial policies, [www.sciencemag.org/authors/science-journals-editorial-policies](http://www.sciencemag.org/authors/science-journals-editorial-policies).

World Wide Web Consortium (2013), PROV-Overview: An overview of the PROV family of documents, [www.w3.org/TR/2013/NOTE-prov-overview-20130430/](http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/).

## Author Information

Zhong Liu ([zhong.liu@nasa.gov](mailto:zhong.liu@nasa.gov)), NASA Goddard Earth Sciences Data (GES) and Information Services Center (DISC), NASA Goddard Space Flight Center (GSFC), Greenbelt, Md.; also at Center for Spatial Information Science and Systems, George Mason University, Fairfax, Va.; Jianwu Wang and Shimei Pan, Department of Information Systems, University of Maryland Baltimore County, Baltimore, Md.; and David Meyer, GES DISC, NASA GSFC, Greenbelt, Md.

## Citation:

Liu, Z.,Wang, J.,Pan, S., and Meyer, D. (2019), Improving reproducibility in Earth science research, *Eos*, 100, <https://doi.org/10.1029/2019EO136216>. Published on 30 October 2019.

Text © 2019. The authors. [CC BY 3.0](#)  
Except where otherwise noted, images are subject to copyright. Any reuse without express permission from the copyright owner is prohibited.

1 Comment

1

Login ▾

G

Join the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS 

?

Name

♥ 1

Share

Best

Newest

Oldest

deepseadawn

4 years ago

Thanks for this great article and sharing a parallel initiative from Arizona State University's Spatial Analysis Research Center <https://sgsup.asu.edu/sparc...>

o

o

Reply

●

Share >

—

🚩

Subscribe

Privacy

Do Not Sell My Data

© 2024 American Geophysical Union. All rights reserved.