Data Reuse and Reproducibility in Earth System Science: A Survey of Current Practices, Barriers, and Expectations

An Yan, Caihong Huang, Carole L. Palmer
yanan15@uw.edu, huangc34@uw.edu, clpalmer@uw.edu
Information School, University of Washington, Seattle, WA, 98195, USA

1. Introduction

As Earth System Science (ESS) becomes more data-intensive, collaborative, and interdisciplinary, data reuse and reproducible research are increasing in importance. This survey study examined the perspectives and practices of ESS researchers on data reuse and reproducibility to inform how to advance future data services in the field.

The findings indicate a strong data sharing culture in ESS with high levels of reuse and commitment to open science. Data reuse and reproducibility of research would benefit most from better documentation and sharing of methods and research processes, and targeted improvements in data services and tools.

2. Survey Design

24 questions in 6 sections

Demographics
- Academic rank, work location
- Research areas

Data Reuse
- Experiences reusing and sharing data
- Purposes for reusing data

Reproducing Published Work
- Experiences reproducing other's studies (barriers, reasons for failures, etc.)

Making Your Research Reproducible
- Practices making research reproducible.
- Perspectives on reproducibility

Possible Remedies
- Reward structure and
- Data service improvement

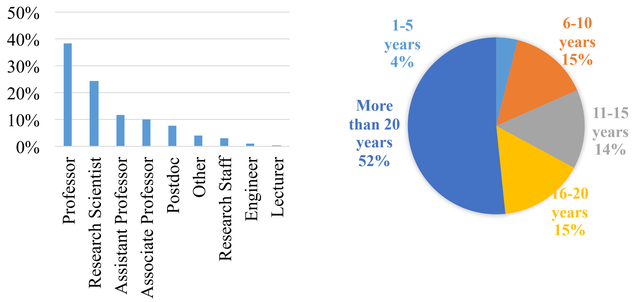
Challenges in Earth System Science
- Issues unique to ESS

References:
[1] Tenopir, Carol, et al. "Data sharing by scientists: Practices and perceptions." PLOS ONE 6.6 (2011).
[2] American Society for Cell Biology (ASCB). "How can scientists enhance rigor in conducting basic research and reporting research results? A white paper from the American Society for Cell Biology." (2015).

This work builds on results from the Site-based Data Curation project (https://datalab.ischool.uw.edu/projects/site-based-data-curation), funded by IMLS National Leadership Grant LG-06-12-0706-12.

3. Survey Administration and Response

2416 ESS researchers were identified based on information from U.S. institutional websites. Invitations to participate in the online survey were distributed through email in two waves, in spring and summer of 2018. Of 225 completed surveys, analysis was conducted on data from **207 respondents from 126 universities and research centers** identified themselves as ESS researchers. The **9.3%** response rate is consistent with comparable online surveys conducted by Tenopir (2011) and ASCB (2015).



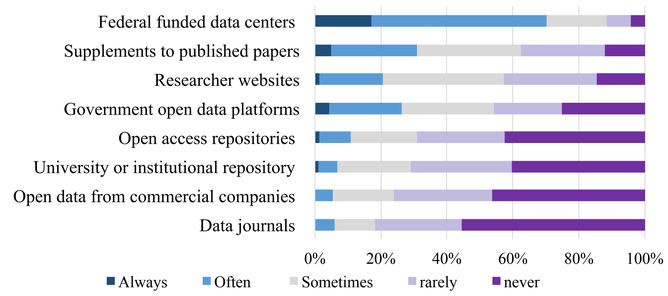
Profile of respondents: position; number of years in the field

4. Results: Sharing and Reuse Practices

A. Data Reuse

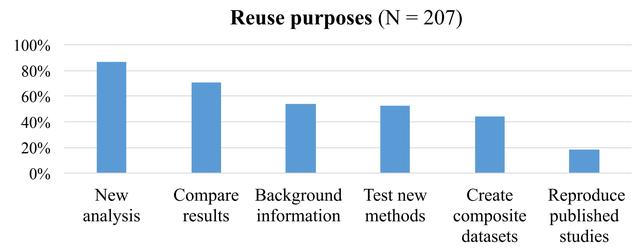
- **96.6% share data; 99% reuse data** generated by others.

How often do you access data produced by others from the following types of sources? (N = 205)



- Main purposes for reusing data:
  87% conduct new analysis, 70.4% compare results, only **18.5% reproduce published studies.**

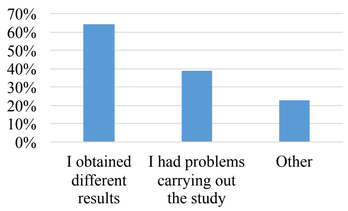
Reuse purposes (N = 207)



4. Results: Practices and Perspectives

B. Research Reproducibility

If you have tried to reproduce a study, how successful have you been? (N = 95)



What was the nature of the failure(s)? (N = 70)

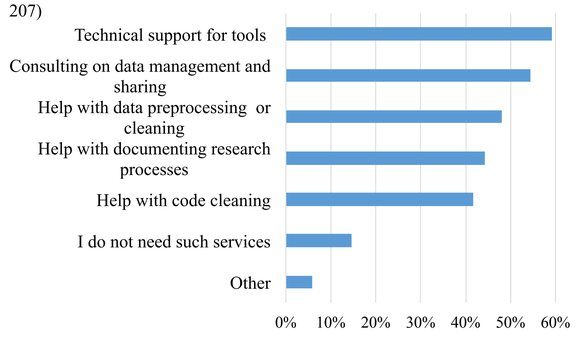
- **The top reproducibility problem**:
  not enough detail in the published paper on how study was conducted - **85.2%**

"Methods in journals are inadequate. There are often missing or unclear details. Reviewers often focus on the overall method and not the details so they get ignored, left out, or errors are not caught before publications. Also, many of the important details are exiled to the supplemental material which receive little if any peer review."

- **96.6% of respondents document their research**, using
  - Word processors or text editors - 83%
  - Handwritten notebooks or journals - 59.9%
  - Code comments - 52.2%

C. Data Services

What kinds of support would you take advantage of, if provided by information professionals at your institution? (N = 207)



"We need staff specifically dedicated to documenting, cleaning, organizing code to facilitate reproducibility elsewhere. When it's time for us to do this, we already have another project we have to work on."

D. Expectations

How much do you agree with the following statements? (N = 207)

Statement	Strongly agree	Agree	Not sure	Disagree	Strongly disagree	Mean
Code and data should be considered important research products.	41.1%	50.2%	6.3%	2.4%	0%	4.3
It should be easier for researchers to publish negative results.	28.0%	50.7%	17.9%	2.4%	1.0%	4.02
Researchers in ESS should have training on reproducibility.	15.0%	53.1%	21.3%	9.2%	1.4%	3.71
Publishers should assess reproducibility as part of the peer review process.	15.5%	47.3%	21.3%	14.5%	1.4%	3.61
Funding agencies should include reproducibility as a criterion for funding.	12.6%	37.7%	29.0%	18.4%	2.4%	3.4
There are adequate expectations for reproducibility in ESS.	2.9%	37.2%	32.9%	25.1%	1.9%	3.14
Researchers should ensure reproducibility of their studies.	48.3 %	44.0%	5.8%	1.4%	0.5%	4.38
Researchers need assistance with reproducible research.	15.9%	40.1%	28.5%	13.0%	2.4%	3.54
Researchers should focus on scientific discovery and minimize the time spent on reproducibility issues.	5.3%	21.3%	17.9%	43.5%	12.1%	2.64

E. Challenges

Responses to open-ended questions suggest:

- Lack of a standard for reproducible research in ESS.
- Reuse is more achievable goal than reproducibility, since
  - complexity of many ESS endeavors makes it difficult to reproduce methods,
  - reproducibility is not applicable to certain types of studies, such as field experiments.
- Large scale of data is growing challenge for reproducibility and reuse.
- Current levels of documentation and descriptions of research methods in published papers are insufficient.
- Some data services require greater technical and scientific expertise.

Conclusion and contributions: Data reuse and sharing are prevalent among ESS researchers, and they are vital for generating new and comparative analyses, but also for gathering background information, testing new methods, and creating composite datasets. Reproducibility is limited by the complexity of ESS research and the nature of field work, but data reuse and research reproducibility would both benefit greatly from improved practices and expectations for documenting data and methods, as well as specialized support, tools, and services for data management and curation. This work laid the groundwork for a more comprehensive and detailed study of how best to advance data services for ESS.