Model Data RCN - Workshop 3 (July 25-27, 2022) Summary Report

PROJECT CONTACTS
Gretchen Mullendore - National Center for Atmospheric Research
Matt Mayernik - National Center for Atmospheric Research
Doug Schuster - National Center for Atmospheric Research
Jared Marquis - University of North Dakota

BACKGROUND

Much of the research in geosciences, such as projecting future changes in the environment and improving weather and flood forecasting, is conducted using computational models that simulate the Earth's atmosphere, oceans, and land surfaces. 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, recent efforts to standardize data sharing and preservation guidelines within research institutions, professional societies, and academic publishers make clear that the scientific community does not yet know what to do about data produced as output from computational models. Guidance to researchers varies and is often unclear. The simplest solution for replicability would be to "preserve all the data", but simulation data can be prohibitively large, particularly in a field like atmospheric science. The massive size of the simulation outputs, as well as the large computational cost to produce these outputs, makes this not only a problem of replicability, but also a "big data" problem. Discussion across different modeling communities suggests that the answer to "what to do about model data" will look different depending on simulation descriptors. Examples of important simulation descriptors include community commitment, simulation workflow accessibility, simulation output accessibility, research feature replicability, and cost of running simulation workflow vs cost of repository data management services.

The ultimate goal of the EarthCube Research Coordination Network (RCN) project "'*What About Model Data?*' Determining Best Practices for Preservation and Replicability" is to develop guidance for authors, funders, and publishers on what data and software elements of simulation based research need to be preserved and shared to meet community open science requirements and expectations. To achieve this goal, two virtual workshops and one hybrid workshop have been held. The first virtual workshop was held from May 5-7, 2020 to craft a draft rubric based on the simulation descriptors that will help researchers and centers describe their simulation data in consistent terms, so that proper decisions are made regarding preservation and retention. The second virtual workshop was held from Aug 3-5, 2020 to test the draft rubric with participant use cases, discuss what simulation workflow components to preserve and why for the various use cases, and discuss general challenges related to the topic of simulation output preservation. The third (hybrid) workshop was held from July 25-27, 2022, in Grand Forks, North Dakota, at the University of North Dakota to discuss issues related to model software and data preservation and sharing that have emerged from discussions in workshops one and two.

This document reports on outcomes from the 3rd project workshop. This workshop was held over 2.5 days (see agenda). The first day started with an overview of the project so far, including the two prior workshops. This update was followed by plenary presentations on each of the four breakout session topics:

1. Sustainable Curation
2. Determining Lifetime for Simulation Data
3. Incentivising Data and Software Sharing
4. Open and Equitable Science

The first breakout session was held at the end of day 1. The remaining three breakout sessions were held on day 2. Each breakout session was followed by plenary summary and discussion. Project leadership and breakout session leaders stayed after the main meeting to document major findings for each topic. Day 3 started with a presentation of major findings from the breakout sessions, and ended with discussion of these findings and brainstorming next steps for the project and the broader community.

OUTCOMES

Major findings were compiled by breakout session topic. A high level summary of these findings is included below.

High Level Findings from Plenary Presentations and Breakout Sessions

1. **Sustainable Curation**
    a. Software and data management plans need to be well thought out by PIs/creators and elevated in importance by funding agencies (broader impact).
    b. Funding should come from agencies specifically for data/software management needs
    c. Incorporate training for data and software management in standard graduate-level curriculum
2. **Determining Repository Lifetime for Simulation Data**
    a. Simulation data do not need to be preserved indefinitely
    b. Plan and advertise de-accession strategy at the point when data is deposited
    c. Use a defined process to evaluate when simulation data can be purged from a repository
3. **Incentivizing Data and Software Preservation and Sharing**
    a. Showcase open science based research success stories across all sectors
    b. Update promotion and tenure process to support sharing of code and data
    c. Raise the visibility of open science achievements through publisher and societal awards
4. **Equitable Access to Data and Software Curation and Analysis Resources**

a. Provide the resources for under-resourced communities to meet open science expectations
    i. Access to data proximate compute and trusted data/software repositories
    ii. Accessible training and support: "National virtual data curation laboratory"

b. Invest in building relationships that include under-represented partners

NEXT STEPS

Workshop participants discussed ideas for the path forward during the final day of the workshop. Below are some ideas that arose during that discussion:

- A "Virtual Data Curation Laboratory", similar to the National Virtual Biotechnology Lab https://science.osti.gov/nvbl might be of use in supporting the Geosciences research community with data and software curation resources and consulting.
- Publishers and sponsors need to work together to develop consistent open science expectations.
- There is a need to identify "Senior/Advanced career" level champions to promote and model the use of open science practices in their research. This would complement the groundswell for embracing open science practices amongst early career researchers.
- Hackathons to update model codes to improve documentation, and build usage templates could be of use to reduce the barriers of entry to effectively sharing well-documented workflows.
- Rubric usability should be improved. Perhaps the users could answer rubric questions through a web survey view instead of using the spreadsheet? A sponsor would need to be found to support and maintain this type of interface.
- Emphasize the importance of documentation in the rubric instructions and use case text. Documentation should describe the comprehensive simulation experiment workflow and resultant data. A documentation strategy should be developed as part of the research project plan.

The project PIs plan to continue to engage community stakeholders through targeted meetings and engagements at disciplinary meetings, and publish on the overall outcomes of the RCN in a peer reviewed journal during the next year. Specifically a town hall to solicit community feedback on "sustainable curation" and "equitable access to data and software curation resources" will be held at the AMS 2023 annual meeting. The PIs also plan to continue engagement with publishers; representatives from AMS and AGU Publishers have already been involved in the project in various ways (steering committee, workshop presentations and participation). Finally, a sustainable path forward to maintain, socialize, and further develop RCN project outputs will be determined.

LINKS TO ADDITIONAL MATERIALS

- Project Website
- Workshop Agenda
- Workshop 3 General Notes Document
- List of Workshop Participants
- Plenary Presentations