Model Data RCN - Workshop 1 (May 5-7, 2020) Summary Report

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

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 reproducible workflows are needed for computational modeling. Open and reproducible 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 archiving guidelines within research institutions, professional societies, and academic publishers make clear that the scientific community does not know what to do about data produced as output from computational models. To date, the rule for reproducibility is to “save all the data”, but model data can be prohibitively large, particularly in a field like atmospheric science. The massive size of the model outputs, as well as the large computational cost to produce these outputs, makes this not only a problem of reproducibility, 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 model descriptors. Examples of important model descriptors include reproducibility, storage vs. computational costs, and accessibility to the community.

The ultimate goal of the EarthCube Research Coordination Network (RCN) project “‘What About Model Data?’ Determining Best Practices for Archiving and Reproducibility” is to provide model data best practices to the community, including publishers and funding agencies. To achieve this goal, a virtual workshop was held May 5-7, 2020, to craft a draft rubric based on the model descriptors that will help researchers and centers describe their model data in consistent terms so that proper decisions are made regarding archiving and retention.

The virtual workshop included three three-hour time blocks on successive days. The first day featured a number of presentations by experts in various aspects of geoscience modeling. The second and third day consisted of breakout discussions with the following goals:

- Breakout session #1 - Participants developed lists of model descriptors with definitions. Following this session, the breakout leaders combined the separate descriptor lists into one large list of model descriptors.
- Breakout session #2 - Participants refined the full descriptor list by combining, refining, and adding new descriptors. In addition, participants started filling in class definitions. The focus was on edge cases in order to to describe the range of possibilities for a given descriptor, e.g. in what cases would one save all of the model output vs. in what cases would one save none of the model output.

After the workshop concluded, the project PIs compiled the descriptor lists, definitions, and classes from the second breakout sessions, and then filtered the lists to eliminate redundancies and descriptors that were not well defined. Although several possible rubric uses were discussed during the workshop, this initial effort was focused as follows: “A rubric to be used to assist a researcher in determining what data or software should be deposited in a FAIR aligned repository to communicate knowledge.” The filtered list of descriptors resulted in the Rubric Version #1, which is described more below. The PIs also compiled notes from all of the breakout discussions into one lengthy notes document. This compiled notes document is also linked below.

OUTCOMES

The model descriptors that came out of the breakout discussions were grouped into a set of themes. Some themes contained many individual descriptors in the final list, while some had a small number.

- Accessibility - This theme includes a range of descriptors related to accessibility of models, model input and output data, the software components, analysis tools, and the system/platforms on which the models are run.
- Community Commitment - The descriptors in this theme touch on specific kinds of modeling projects where the outputs are important to larger communities, such as MIPs or benchmark studies.
- Cost - This theme covers various cost considerations for computational model-based research, including computing, storage, and curation costs.
- Reproducibility - This theme includes descriptors that focus on different aspects of reproducibility, as that term relates to model-based science.

Additional themes were included in earlier versions of the rubric, but were eliminated from the Rubric Version #1 during the post-workshop activities of filtering and compiling descriptors. These themes continue to be important in discussions of model data archiving and reproducibility, but were deemed as either applicable to other rubric users (i.e., not individual researchers communicating knowledge) and/or too subjective for practical use.

- Provenance - This theme covers the metadata, documentation, and provenance information needed to understand and use models and their output.
- Longevity - This theme includes descriptors that relate to long-term considerations, including the sustainability of the model software and data formats.

- Value Judgement - The descriptors in this theme cover a range of issues that are involved in judging the value of particular model outputs, including aspects of temporal and spatial coverage, error correction, and potential utility of the outputs for new studies.
- Versioning - This theme covered issues related to the versioning of model code, including the documentation of the version changes, the accessibility and usability of the various model versions, and the difference in output among different model versions.

LINK TO DRAFT RUBRIC

The compiled and filtered rubric - available as PDF or xlsx - contains the summarized and prioritized list of model descriptors. To repeat from above, the motivating use case for this version of the rubric was: “A rubric to be used to assist a researcher in determining what data or software should be deposited in a FAIR aligned repository to communicate knowledge.”

NEXT STEPS

The next workshop for this project is being planned as another virtual workshop to take place August 3-5, 2020. The focus of the next workshop will be to assess the rubric with a range of test cases that represent different kinds of models and model experiments. These test cases will help to refine the model descriptors contained in the rubric, and the guidance on how to use the rubric.

Between workshops, the draft rubric will be presented at the EarthCube Annual Meeting and the Community Earth System Model (CESM) Workshop.

LINKS TO ADDITIONAL MATERIALS

- Project Website
- Workshop Agenda
- Workshop Welcome Video
- Compiled notes document - includes all notes from all groups, somewhat organized by the overarching themes noted above.
- Full descriptor list - includes all descriptors created/described by all breakouts. In many cases the same descriptors are listed multiple times, to illustrate different definitions and class distinctions from the different breakout groups.
- List of workshop registrants
- Plenary Presentations