Determining Best Practices for Archiving and Reproducibility of Model Data

Gretchen Mullendore, U. North Dakota
Matthew Mayernik, NCAR
Doug Schuster, NCAR
(presented by Gary Strand, NCAR)

NSF Awards #1929773, #1929757

Community involvement

- Project funded in October, 2019
  - No results yet, so why this talk?
- EarthCube Research Coordination Network (RCN)
  - Community collaboration on data science needs of the geosciences
  - We're recruiting!
- Today's outline:
  - What's the problem?
  - What's the plan to tackle this problem?
  - How can you sign up?

The rise of data management

- Atmospheric sciences has been ahead of other communities in acknowledging the need for thoughtful data management
  - Big data issues in storage and analysis
  - Need to maintain observations permanently
  - Data sharing among many users; need for data standards
  - Push for open access and reproducibility
- Advances in data management have led to data management requirements from consortia, funding agencies, and publishers

AGU Publications Data Policy (2016)

- "...all data necessary to understand, evaluate, replicate, and build upon the reported research must be made available and accessible whenever possible. For the purposes of this policy, data include, but are not limited to, the following:
- Data used to generate, or be displayed in, figures, graphs, plots, videos, animations, or tables in a paper.
- New protocols or methods used to generate the data in a paper.
- New code/computer software used to generate results or analyses reported in the paper.
- Derived data products reported or described in a paper."

AGU: KEEP ALL THE DATA

AMS Data Archiving and Citation Guidelines (2019)

"For authors, this means that at initial submission of the manuscript, they must confirm that their data are archived and cited/referenced properly… data includes environmental predictions generated by numerical models, and data products derived from integrations of observational and model-generated sources."

"Authors are expected to direct all core research outputs to FAIR-aligned repositories, following the FAIR principles."

AMS: DO WHAT FAIR SAYS

Enabling **FAIR** Data FAQs (COPDESS, 2019)

Q 2.9. My data are mostly model output or are derived summary figures based on numerous model runs and outputs. The collective output could easily exceed 1 TB. What am I required to deposit?

A. …If you are part of a modeling center that has a standard archiving plan for models and run output, please follow that plan (such centers are an acceptable repository). In general, the most important information to provide on models is the code (and version used) and unique configurations, any input parameters, run files, and a description of the overall run environment and parameter space tested. Representative output can also be deposited, ideally key data that form the figures included in the paper.

Q 2.13. My data are huge, terabytes or petabytes in size. How can I share them?

A. At the time of this writing, data of this size are best stored on-site, as the data are too large to transmit and exceed the limits of most domain or general repositories. Check with your repository on any size limits. If you are regularly collecting or producing this much data, you should check with your funder, facility, or institution for an on-site archiving plan. If these data are model output, please see Q 2.9.

FAIR: KEEP SOME OR ALL DATA

AND ASK SOMEONE ELSE

What to do about model data?

- Models: We know the answer is not “save all the data”
- What do we do?
- Bring together a diverse group of modeling experts to:
  1. Develop a comprehensive list of model descriptors
  2. Use model descriptors to build rubric for model rankings
  3. Refine rubric with extensive set of use cases
  4. Disseminate best practices document to broader community

Model Descriptors and Ranking Rubric

SAVE OUTPUT ⟷ SAVE SETUP

Model Descriptor	Rank 1	Rank 2	Rank 3
Value to Scientific Community	Of use across many research disciplines	Of use to multiple researchers, but only a single discipline	Tailored to a particular research question with minimal value to other studies
Reproducibility	Would be difficult to reproduce due to nonlinearity of phenomena being studied	Would be difficult to reproduce some details, but general findings are robust	No issues with reproducibility (could be due to study subject or to model packaging, e.g. containerization)
Computational Cost	High computational cost and can only be produced with specialized platforms	Moderate computational cost, but access to needed platforms straightforward	Small computational cost with no special platform needs

Other descriptor examples:

- Reproducibility: bitwise versus physical
- Results longevity
- Model version longevity
- Storage costs

Many models not clear cut; will look for model categories with use cases

Use case examples

- Idealized parameterization study
  - Easy to rerun, robust model differences, keep initializations only
- Climate model intercomparison project
  - Widely used, computationally expensive, large storage costs, keep outputs on a agreed upon standard output grid
- Ensemble modeling system of nonlinear phenomenon
  - Individual members have low reproducibility, but ensemble statistics robust, computationally expensive, large storage costs, ?
  - Depends on value to community and more careful weighing of computational costs versus storage costs/access request frequency

Year 1: Workshops

- Workshop participant travel funded by NSF
- Workshop #1 - Boulder, Colorado, May 5-8, 2020
  - Brainstorm and prioritize model descriptors.
  - First draft of model rubric.
- Workshop #2 - Grand Forks, North Dakota, Aug. 3-6, 2020
  - Refine rubric with use case testing.
  - First draft of recommendations for different model categories
- Participants:
  - Experienced modelers from a wide range of disciplines
  - Inclusion of advanced graduate students and early career scientists

Year 2: Community Feedback

- Presentations/Town Hall discussion at conferences
  - e.g., EarthCube Annual Meeting, AMS Annual Meeting, AGU Fall Meeting, Earth Science Information Partners (ESIP), WRF Workshop, CESM Workshop
- Community document for comment

- Goal: Provide a best practices document to community
  - Will not solve all our model data archiving and reproducibility problems!
  - Will provide a common framework for discussing model data needs for experts *and non-experts*

Steering Committee

- **Adam Clark**, NOAA/University of Oklahoma
- **Gilbert Compo**, Cooperative Institute for Research in Environmental Sciences (CIRES), UC Boulder
- **Laura Condon**, University of Arizona, Hydrology and Atmospheric Sciences
- **Gokhan Danabasoglu**, NCAR, Climate and Global Dynamics Laboratory
- **Josh Hacker**, Jupiter
- **Michael A. Friedman**, American Meteorological Society (AMS)
- **Matthew Mayernik**, NCAR, Library, co-PI
- **Gretchen Mullendore**, University of North Dakota, Atmospheric Sciences, co-PI
- **Douglas Schuster**, NCAR, Computational & Information System Laboratory, co-PI
- **Gary Strand**, NCAR, Climate and Global Dynamics Laboratory

Student members:

- **Jared Marquis**, University of North Dakota, Atmospheric Sciences
- **Elisa Murillo**, University of Oklahoma, School of Meteorology

How to Get Involved!

- *Virtual Workshop #1 - Boulder, Colorado, May 5-7, 2020*
  - Agenda
  - Workshop Welcome Video
  - May 5 Plenary Presentations
  - Workshop Outputs
- Virtual Workshop #2 - Grand Forks, North Dakota, Aug. 3-6, 2020
- **If interested in a workshop invitation**, please email Doug Schuster (schuster@ucar.edu)
  - A brief statement on why you are interested in participating
  - Background on your experience with computational models and data

https://modeldatarcn.github.io