"What About Model Data?"
Best Practices for Preservation and Replicability

User Guidance

Document Purpose: Information for understanding and using rubric and use case materials

Current Rubric Worksheet:
PDF version:
https://modeldatarcn.github.io/rubrics-worksheets/Descriptor-classifications-worksheet-v2.0.pdf
Microsoft Excel Version:
https://modeldatarcn.github.io/rubrics-worksheets/Descriptor-classifications-worksheet-v2.0.xlsx

Document Organization:
1. Project Overview
2. Data Production versus Knowledge Production
3. Rubric Instructions
4. Guidance for Data Production Projects
5. Use Case Instructions -What to Preserve and Share?
6. Use Case Examples -What to Preserve and Share?

1. Project Overview

There is strong agreement across the sciences that replicable[1] 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 know what to do about data produced as output from computational models. To date, the rule for replicability is 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. The ultimate goal of the EarthCube Research Coordination Network (RCN) project “‘What About Model Data?’ Determining Best Practices for Preservation and Replicability” is to provide simulation data management best practices to the community, including publishers and funding agencies.

2. Data Production versus Knowledge Production

The majority of research involving simulations is knowledge production not data production[2]. In other words, the primary goal of most projects involving computer simulations is to increase scientific knowledge and the simulations are used as a tool to that end. Data production projects (e.g., CMIP), in contrast, are motivated by scientific questions, but the primary goal is to provide a dataset that multiple users can access to investigate those scientific questions. While most researchers that produce simulation output would welcome more use of their output products, and many end users would welcome more data availability, the reality is that we are producing far more simulation output than can be reasonably stored in repositories. Knowledge production research should preserve minimal output in repositories. Further guidance in determining whether a given project is knowledge or data production is provided in the Rubric Instructions.

3. Rubric Instructions

The rubric is built to help researchers make decisions about what simulation output needs to be shared via a repository, i.e. made accessible and preserved a sufficient time to satisfy the requirements of publishers and funding agencies. Ultimately, these decisions are based on the goal of all community members (researchers, publishers, end users) to communicate knowledge in a sustainable way. Note this rubric is not meant to dictate what a researcher or research group keeps on their own local storage.

The rubric is a list of simulation/experiment descriptors, organized into themes. To use the rubric, consider a specific simulation workflow and review all of the descriptors, selecting a Class (1, 2, or 3) that best fits your workflow for each descriptor. If you find that your workflow could be a Class 1 or 3 based on which descriptor you are looking at, you can separate the workflow into logical sections and then classify each section separately.

Once you have selected classes for each descriptor, multiply classifications of 2 or 3 by the recommended weighting shown in the “suggested weighting” column. Certain researchers may have compelling reasons to adjust the weighting based on the goals of a specific project, but testing thus far has shown these recommended weights to be useful for most projects.

Finally, the total weighted score will indicate what to deposit into a repository: few outputs, some outputs, or the majority of outputs (see score ranges at the bottom of the rubric). Further guidance on specifically what outputs to save are provided in the Use Case Example document. See instructions for using that document in section 5 below.

4. Guidance for Data Production Projects

If your workflow scored high in the first theme area of the rubric, Community Commitment, your research may be a data production project, regardless of your total score. Data production projects are designed to have significant products preserved with wide accessibility. These projects hopefully included an appropriate budget to support anticipated data preservation and community data access needs.

Although outputs preserved are often broadly dictated by the research proposal for data production research, the rubric and use case examples may still be useful. For example, additional decision-making may be needed regarding what specific outputs should be preserved and shared. This project can also be helpful when communicating with end users about why not all the outputs can be made available.

5. Use Case Instructions -What to Preserve and Share?

The purpose of the Use Case Example document is to give examples of other projects in a particular rubric score range. Specifically, what data are being preserved and why? As all projects are unique, there is no one solution for all projects, even when rubric scores are identical. Instead, the Use Case document is meant to assist a researcher in making thoughtful and informed decisions on what data preservation is necessary for replicable science. Please note that the Use Case Example document is a first draft version.

While each project is unique, certain data should be preserved and shared for all projects. The following components should be preserved and shared unless they are already publicly available from another provider:

- simulation code
- initialization data
- simulation setup (e.g., parameterization selection)
- pre-processing code
- post-processing code

Note that while we encourage full preservation and sharing for all above categories, several challenges have been identified in workshop discussions, including proprietary data/code and giant non-accessible datasets (e.g., operational forecasts). Examples of these are included in the Use Cases. The challenge of sharing code also includes the continuing concerns that code is not cited properly, such that sharing code can be detrimental to researchers getting credit for their work. This last issue is a challenge, but not a barrier, and we must continue to work towards better code/data citation practices (and recognition of such in job evaluations) to achieve our goals of open, efficient, and replicable science.

With the above components saved for all projects, the Use Case Example document can then be used to make decisions on preserving simulation output. As discussed in section 2 of the instruc, most projects are Knowledge Production research, and as such should be saving little to no raw simulation data in repositories and focus on 2-D derived fields that help communicate to future researchers the environmental state or other information important for building similar studies in the future. Particularly for highly nonlinear case studies, the goal is not exact reproducibility, but rather enough output to understand the environmental state that forced, and the impacts of, the features being investigated.

6. Use Case Examples -What to Preserve and Share?

Three reference use cases on “what to preserve and share” were compiled through project workshops. Please find a summary of each use case below.

Use Case 1, Knowledge Production - *Preserve few simulation workflow outputs*

- Example Use Case Description
    - Semi-idealized WRF-ARW-based numerical simulations of tropical cyclones over land. Involves some code modifications, primarily to the land-surface model (e.g., to fully disable radiative transfer and/or to partially or fully disable surface latent and sensible heat fluxes). Involves extensive initial-condition modifications to both atmospheric and land-surface parameters, primarily to homogenize the atmospheric and land-surface states.
- What should be preserved and shared?
    - Input - initialized data from GFS output, took a sounding and interpolated it to the model grid. If possible point to the NWP center or that produced this data or responsible long-term institutional archive such.
    - Model configuration - namelist file
    - Code used for interpolation and sounding data
    - Model code - changes to NOAH LSM to fully disable radiative transfer and/or to partially or fully disable surface latent and sensible heat fluxes - want to tar up the whole model (including WRF) to make it easier for reuse
    - Raw output **-None**
        - In weather forecasting, don't keep raw 3-D output, keep 2-D diagnostic fields instead
    - Processed output
        - preserve processed hourly averaged files (2-D derived fields)
        - Optional: use GRIB to preserve diagnostic fields (share GRIB table with it as important metadata); has some advantages for disk resources when most of the field is 0 (e.g. precipitation)
    - Processing code
        - Making available custom post-processing code. Link to open source post processing tools where these are available.
- Why should it be preserved and shared?
    - Sharing model code modifications back to the community as appropriate is a good practice.
    - Don't necessarily need to share/document every parameter change
    - Documentation is important for code. Use of diff command to track changes, and describe what was changed and why (at minimum?), share tar-ball with these comments
    - Benchmarking could be made possible by 2-D diagnostic fields, to capture environment
    - Feature reproducibility is a problem in really nonlinear case - may need to do containerization, etc to be able to capture a more granular level of information - but still may not need the raw output, and leave feature reproducibility to the side

Use Case 2, Knowledge Production- *Preserve selected simulation workflow outputs*

- Example Use Case Description

    - Warn-on-Forecast - Short-term (0-6hr) convection-allowing (3-km) ensemble forecast system aimed at severe weather prediction. Limited area (900x900km?)
    - 18 WRF-ARW members, 15-min data assimilation frequency (incl. radar and satellite)
    - Forecasts every 30 min with probabilistic outputs (web interface)
    - About 100 TB raw data (netCDF) for spring season; preserved data reduced to about 1 TB per case (about 25TB total)
- What should be preserved and shared?
    - Initiation and assimilation data - Base fields and boundary conditions from HRRRE (HRRR Ensemble). If possible point to the NWP center or that produced this data or responsible long-term institutional archive.
    - Codes: Model, pre and post-processing, DA (GSI)
    - Scripts for running each version of Warn-on-Forecast
    - Should be developing and saving detailed documentation to run the code.
    - Raw simulation output should NOT be saved (files are too large). Important fields and storm diagnostics extracted with post-processing software are saved, which is only a fraction of the size of the raw output.
    - Visualizations and web images to easily inspect past cases
- Why should it be preserved and shared?
    - To test changes to model/DA/preprocessing - if WRF input and WRF boundary condition files are saved, it is easy to replicate the simulation runs and produce the raw output if needed.
    - The model source code absolutely should be preserved and shared because it's been heavily modified from publicly available versions.

<u>Use Case 3, Data Production</u> - ***Preserve the majority of simulation workflow outputs***

- Example Use Case Description
    - Using the CMAC model to study ammonia in the atmosphere. Running WRF on CONUS, and perturbe it, so there are multiple versions of the output. They are huge files, and three copies due to the perturbation runs.
    - This is a NASA funded project. NASA wants others to use what is created via this project.
    - What should be preserved and shared?
        - Output data preserved with all parameters
        - Notes on how it was produced because it probably won't be possible to reproduce
        - Model code and relevant scripts
        - Data archive consideration - If a project needs to preserve and share very large data volumes (10s - 100s of TBs), remote access becomes more difficult. The project may need special services to support access (e.g. subsetting for data volume reduction or community accessible, data proximate compute resources). The project may also need different approaches to store and access data based on how many users the data will receive.

- Why should it be preserved and shared?
  - Takes a long time to compute and post-process
  - Output data are being generated for any user
  - Planning to develop an interface to allow people to select data based on geographic region, to reduce download volumes.
  - Important distinction between development runs and production runs
    - Good software engineering and documentation should enable rerunning old versions if necessary.
  - May not have control over hardware, which might change and cause difficulties in recreating exact output
  - Difficulty in regenerating outputs, either by yourself or the potential users, lean toward keeping the outputs.

[1]National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

[2]Baker, K. S., and Mayernik, M. S. 2020. Disentangling knowledge production and data production. *Ecosphere* 11(7):e03191. 10.1002/ecs2.3191

[3]See https://zenodo.org/record/3482769#.X-zbd-lKifU and https://zenodo.org/record/3479199#.X-zbU-lKifU