"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://gdex.ucar.edu/dataset/14_schuster/file/Descriptor-classifications-worksheet-v2.0.pdf
Microsoft Excel Version:
https://gdex.ucar.edu/dataset/14_schuster/file/Descriptor-classifications-worksheet-v2.0.xlsx
Example Interactive Rubric:
https://modeldatarcn.github.io/rubrics-worksheets/rubric-example.html

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 have clarity on 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 develop guidance on what data and software elements of simulation based research need to be preserved and shared to meet community open science expectations, including publishers and funding agencies. It is recommended that researchers work through the rubric when the project is being formulated, to estimate and include any necessary data management costs in the proposal budget at that time.

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 and accompanying use case examples are intended to assist researchers in determining what simulation output needs to be shared through a trusted community repository to communicate knowledge, thus satisfying the requirements of publishers and funding agencies. Ultimately, these decisions are based on the goal of all community members (e.g., Researchers, Publishers, Research Consumers) to transparently communicate knowledge in a sustainable way. Note this rubric is not intended to dictate what a researcher or research group keeps on their own local storage.

The rubric is composed of a list of simulation/experiment descriptors, which are organized into the following themes:

- **Community Commitment**
  - Is it anticipated that your simulation workflow outputs will have broad community impact and downstream reuse?
- **Research Workflow Accessibility**
  - Would it be reasonable to expect others in your academic discipline to rerun your full simulation workflow?
- **Data Accessibility**
  - Would it be reasonable to expect others to access and use simulation workflow outputs?
- **Research Feature Replicability**
  - Are physical features generated by a simulation replicable?
- **Cost**
  - Is it more costly to re-run a full simulation workflow or preserve model output products in a trusted repository?

Themes are broken out as individual components in the rubric, and there can be multiple components describing a theme. This allows users to view the contributions to the total rubric score from each rubric theme component. One example is the ”Cost” theme. There are two components in the rubric that examine cost: 1) cost of running the simulation workflow and 2)

repository data management services cost. Through these two cost components, the rubric is attempting to determine if it would be more cost effective to have research consumers with domain knowledge re-run the full simulation workflow (Class 1 -Preserve Few Outputs) or to have research consumers access simulation output products through a trusted community repository (Class 3 -Preserve Most Outputs).

To use the rubric, consider a specific simulation workflow, review all of the descriptors, and select a Class (1, 2, or 3) that best fits the characteristics of the simulation workflow for each descriptor. If it is found that a workflow could be a Class 1 or 3 based on which descriptor is being reviewed, separate the simulation workflow into logical sections and then classify each section separately.

Once classes have been selected for each descriptor and entered into the scoring column for a descriptor, multiply scores of 2 or 3 by the recommended weighting shown in the “suggested weighting” column of the rubric. 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. Further guidance on specifically what outputs to save are provided in the Use Case Examples. 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.

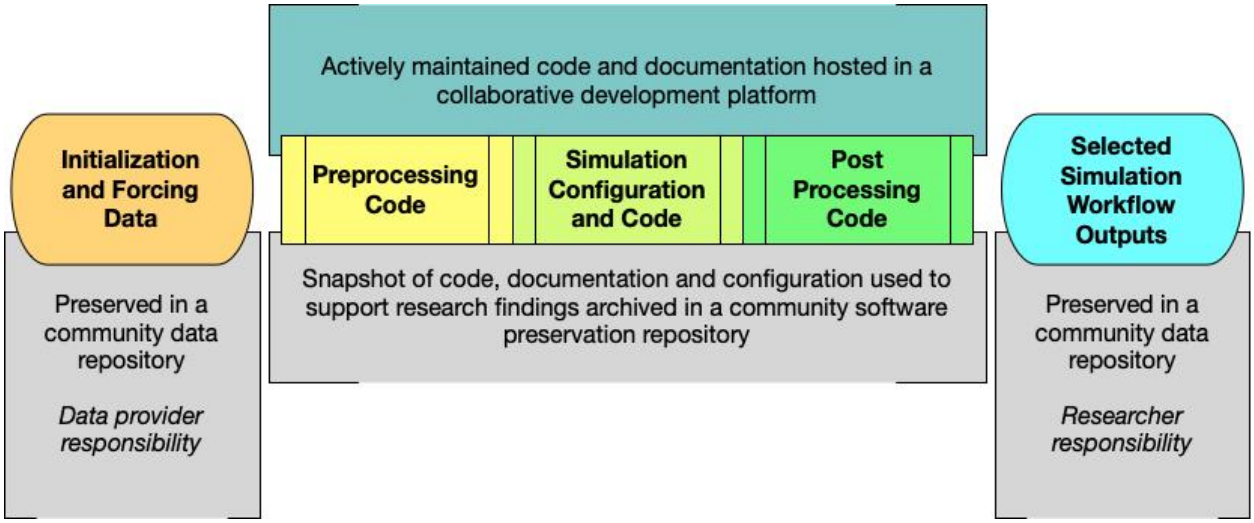
Although outputs preserved are often broadly dictated by the research proposal for data production research, the rubric and use case examples may still be useful. For example, additional decision-making may be needed regarding what specific outputs should be preserved and shared. This project can also 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.

While each project is unique, certain data and software components should be preserved and shared for all projects. The following components should be preserved and shared[4] unless they are already publicly available from another provider:

- Initialization and forcing data
- preprocessing code
- simulation configuration and code (e.g., parameterization selection)
- 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. One challenge of sharing software source code also includes the continuing concern that software 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 preserved and shared 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 instructions, 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. Following the examples, a Use Case worksheet template is available for individual use.

<u>Use Case 1</u> *-Preserve Few Simulation Workflow Outputs*

- Use Case Description
    - High-level overview of the use case
        - *Semi-idealized WRF-ARW-based numerical simulations of tropical cyclones over land.*
    - Science goals and basic workflow
        - Science goals: *Test the sensitivity of overland tropical cyclone intensity change (in non- to weakly baroclinic environments) to soil characteristics (type, temperature, moisture - and thus heat capacity, thermal conductivity, etc.), land-surface physics (latent and sensible heat fluxes underneath the simulated cyclone vs. in the external inflow environment), and large-scale atmospheric moisture. More generally, the goal of this study is to reconcile differences in existing hypotheses posed to explain why some tropical cyclones are able to maintain or increase their intensity well inland.*
        - Workflow: *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.*
- For all projects: what to preserve (see sections below)
    - simulation code
    - initialization data
    - simulation setup (e.g., parameterization selection)
    - pre-processing code
    - post-processing code
- What use-case specific additional materials should be preserved and shared? [the lists below are possibilities]
    - Data
        - Inputs to model
            - Description
                - *GFS output: took a sounding and interpolated it to the model grid. If possible point to the NWP center that produced this data or responsible long-term institutional archive.*
            - *Total data volume preserved in a repository maintained by an outside data provider (e.g. NCEI): 1 MB (only need input soundings; text files)*
        - Raw model output
            - *Total data volume not preserved in a repository? (might be retained on PI's local working storage): 1-5 TB for*

ensembles/suites of convection-allowing simulations (keep for several years)

- Processed model output
  - Description
    - *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)*
  - *Total data volume preserved in a repository by the PI: Unknown at this time (PIs not used to this approach)*

- Software (see above sections)
  - Model configuration
    - *Yes, namelist file*
  - Preprocessing code
    - *Yes, code used for interpolation of sounding data*
  - Model code
    - *Yes, 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 (or could store as container image)*
  - Postprocessing code
    - *Yes, making available custom post-processing code. Link to open source post processing tools where these are available.*
- Other:
  - Documentation [everything should have this, but maybe there are special kinds of documentation produced for particular use cases]
    - *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. Also share resulting publications as additional motivation for why changes were made. Zenodo used for sharing the tarball (easy to use, provides DOI for citation and tracking).*
  - Visualizations or images [products intended to be used visually, distinguished from processed output that exists as numerical data]
    -

- Why should these things be preserved and shared?
  - Reasons why the things listed above are important
    - General
      - *Sharing model code modifications back to the community as appropriate is a good practice.*
      - *Benchmarking could be made possible by 2-D diagnostic fields, to capture environment*

        - Note expected/intended audience and what they expect/need
            - Are there specific people who will be using the data downstream?
                - *Most likely users: Colleagues and/or students conducting research in the same area. Need to be able to successfully follow on this analysis. Goal is not exact reproducibility.*
            - Possible/aspirational users?
                -
        - Note any temporal considerations, such as particular products that become more/less useful over time
            -
    - Could refer to individual rubric descriptors in this section - which descriptors are most important/useful to guide the preservation recommendations for each case?
        -

<u>Use Case 2</u> *-Preserve Selected Simulation Workflow Outputs*

- Use Case Description
    - High-level overview of the use case
        - *This use case consists of model data covering 25 days during which the Warn-on-Forecast System (WoFS) was run during Spring of 2020. The WoFS is an on-demand, experimental, convection-allowing (3-km grid-spacing) ensemble forecast system with a rapidly updating data assimilation system aimed at extending lead times for hazardous weather. WoFS runs over movable limited area domains (900 x 900 km) providing forecasts to 6 h, and uses the Advanced Research Weather Research and Forecasting Model (WRF-ARW).*
    - Science goals and basic workflow
        - *The use case is a prototype configuration for WoFS, with the idea that a similar type of system will be transitioned to operations in the NWS in the next few years. The use case serves as a baseline against which skill for future configurations can be assessed. The dataset also provides various forecast fields for testing forecast diagnostics, post-processing, and visualization strategies. The workflow includes pre-processing of observational datasets (e.g., radar, satellite, surface observations, aircraft, etc.), data assimilation, forecast integration, post-processing, and visualization via a web-viewer.*
- For all projects: what to preserve
    - simulation code
    - initialization data
    - simulation setup (e.g., parameterization selection)
    - pre-processing code
    - post-processing code

- What use-case specific additional materials should be preserved and shared? [the lists below are possibilities]
    - Data
        - Inputs to model
            - Description
                - *Observations that are assimilated include Multi-Radar, Multi-Sensor (MRMS) reflectivity, MRMS radial velocity, GOES-16 ABI cloud water path & all sky radiances, NCEP prepbufr, and Oklahoma Mesonet data (when available).*
                - *Background data is derived from (1) the 9-member, 31-h forecasts of the 0600 UTC initialized High-Resolution Rapid Refresh Ensemble (HRRRE), (2) 36-h forecasts of the 1200 UTC initialized HRRRE, and (3) 36-member, 1-h forecasts from the HRRR Data Assimilation System (HRRRDAS).*
                - *Input files produced by the WoFS include, (1) 36-member WRF boundary files (9-member HRRRE forecast duplicated thrice) from 0600 and 1200 UTC HRRRE, (2) 36-member WRF input files (from 36-member HRRRDAS) at initialization time. At every 15-min WoFS cycling period, 36 WRF inputs are produced (following data assimilation), 72 WRF output (prior & posterior) are produced, and a text file is produced with info from each assimilated observation.*
            - *Total data volume preserved in a repository by the PI: 81 TB*
        - Raw model output
            - Description
                - *For each WoFS forecast cycle (every 30 minutes from 1700 to 0300 UTC), pertinent WRF variables are output every 5-minutes for all 18 members in standard WRF netcdf format (i.e., “wrfouts”).*
            - *Total data volume not preserved in a repository (might be retained on PI’s local working storage): 200 TB*
        - Processed model output
            - Description
                - *Important environmental fields and storm diagnostics extracted with post-processing software are saved in the form of “summary files”, which are only a fraction of the size of the raw output.*
            - *Total data volume preserved in a repository by the PI: 45 TB*
    - Software
        - Model configuration
            - *Yes, includes namelists and configuration files*
        - Preprocessing code

            - *Yes, includes WRF pre-processing and GSI data assimilation*
        - Model code
            - *Yes, includes specialized version of WRF-ARW*
        - Postprocessing code
            - *Yes, includes specialized python codes for creating the summary files and computing various diagnostics*
    - Other
        - Documentation [everything should have this, but maybe there are special kinds of documentation produced for particular use cases]
            - *Detailed documentation for running the code has been written and is available.*
        - Visualizations or images [products intended to be used visually, distinguished from processed output that exists as numerical data]
            - *Visualizations and web images are available from the URL: https://wof.nssl.noaa.gov/realtime/.*
- Why should these things be preserved and shared?
    - Reasons why the things listed above are important
        - General
            - *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.*
        - Note expected/intended audience and what they expect/need
            - Are there specific people who will be using the data downstream?
                - *WoFS model developers, scientists developing AI and machine-learning algorithms for post-processing, and forecasters interested in potential utility.*
            - Possible/aspirational users?
                - *Students, emergency managers, etc.*
        - Note any temporal considerations, such as particular products that become more/less useful over time
            -
    - Could refer to individual rubric descriptors in this section - which descriptors are most important/useful to guide the preservation recommendations for each case?
        -

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

- Use Case Description
    - High-level overview of the use case
        - *Using the Community Multiscale Air Quality model (CMAQ, https://www.epa.gov/cmaq) to study ammonia in the atmosphere. They*

run WRF to then run MCIP. This data is then used in CMAQ and SMOKE. CMAQ is perturbed and run with different scenarios to find a difference. This creates lots of data in different stages of the project and resulting in many output files.
        - *This is a NASA funded project. NASA wants others to use what is created via this project.*
    - Science goals and basic workflow
        - *To better estimate the ammonia emissions using NASA Cross-track Infrared Sounder (CrIS) data in conjunction with CMAQ.*
        - *There are multiple stages to this project, including running WRF, MCIP, SMOKE, iterating CMAQ, and processing the CRIS data.*
        - *The final product of this project will be a file with gridded ammonia emissions for North America, which is relatively small compared to what it took to create it.*
- For all projects: what to preserve
    - simulation code
    - initialization data
    - simulation setup (e.g., parameterization selection)
    - pre-processing code
    - post-processing code
- What use-case specific additional materials should be preserved and shared? [the lists below are possibilities]
    - Data
        - Inputs to model
            - Description
                - *None are needed to be preserved because the inputs are from NOAA and NASA data archives, and easily available.*
        - Raw model output
            - Description
                - *For long-term preservation - None*
                    - *Don't need output from intermediary stages once they are past those stages.*
                - *For short-term saving - They go through their version of the rubric for each model at the end of each stage. They delete some data at that point, and save some portions of the interim model output in case there may be some need to revisit it, because it would be easier than having to rerun the model. But they plan to only save it for some period of time before deleting (with the time period depending on potential use).*
        - Processed model output
            - Description
                - *CMAQ emissions profile - Output data (only the ammonia-related variables, ~Gb)*

    - Software
        - Model configuration
            - *Yes*
        - Preprocessing code
            - *Yes*
        - Model code
            - *Yes*
        - Postprocessing code
            - *Yes*
        - *All models and scripts are Dockerized. They will probably keep these forever because it was hard to do.*
    - Other
        - Documentation
            - *Notes on how the model output was produced because it probably won't be possible to reproduce*
            - *Documentation on workflows and docker containers for each stage, e.g. notes on where input data came from, such as NOAA, and the settings/versions used for compiling the models.*
        - Visualizations or images
            -
- Why should these things be preserved and shared?
    - Reasons why the things listed above are important
        - *Output data are being generated for any user*
        - *Takes a long time to compute and post-process the model output*
        - *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 of the model 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.*
    - Could refer to individual rubric descriptors in this section - which descriptors are most important/useful to guide the preservation recommendations for each case?
        - *Cost to store vs cost to rerun including labor hours*
        - *Note: Amy and AER developed a custom version of the rubric for use internally.*

Use Case Template *-Yellow Highlighted text to be completed by rubric user*

- Use Case Description
    - High-level overview of the use case
        -

    - Science goals and basic workflow
        -
- For all projects: what to preserve (see sections below for details)
    - simulation code
    - initialization data
    - simulation setup (e.g., parameterization selection)
    - pre-processing code
    - post-processing code
- What use-case specific additional materials should be preserved and shared? [the lists below are possibilities]
    - Data
        - Inputs to model
            - Description
                -
            - *Total data volume preserved in a repository by the PI*
            - *Total data volume preserved in a repository maintained by an outside data provider (e.g. NCEI)*
            - *Total data volume not preserved in a repository? (might be retained on PI's local working storage)*
        - Raw model output'
            - Description
                -
            - *Total data volume preserved in a repository by the PI*
            - *Total data volume not preserved in a repository? (might be retained on PI's local working storage)*
        - Processed model output
            - Description
                -
            - *Total data volume preserved in a repository by the PI*
            - *Total data volume not preserved in a repository? (might be retained on PI's local working storage)*
    - Software
        - Model configuration
            -
        - Preprocessing code
            -
        - Model code
            -
        - Postprocessing code
            -
    - Other
        - Documentation [everything should have this, but maybe there are special kinds of documentation produced for particular use cases]
            -

        - Visualizations or images [products intended to be used visually, distinguished from processed output that exists as numerical data]
            -
- Why should these things be preserved and shared?
    - General
        -
    - Reasons why the things listed above are important
        - Note expected/intended audience and what they expect/need
            - Are there specific people who will be using the data downstream?
                -
            - Possible/aspirational users?
                -
        - Note any temporal considerations, such as particular products that become more/less useful over time
            -
    - Could refer to individual rubric descriptors in this section - which descriptors are most important/useful to guide the preservation recommendations for each case?
        -

[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

[4]Mullendore GL, Mayernik MS and Schuster DC (2021) Open Science Expectations for Simulation-Based Research. Front. Clim. 3:763420. doi: 10.3389/fclim.2021.763420