Model Data RCN - Workshop 2 (Aug 3-5, 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 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 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. 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 provide simulation data management best practices to the community, including publishers and funding agencies. To achieve this goal, two virtual workshops have been held to kick off the project. 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, which is the focus of this report, 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 2nd virtual workshop included three, three-hour time blocks on successive days. The first day featured a plenary presentation session on reference use cases to test the rubric. Presenters included researchers from a variety of geosciences disciplines. This was followed by a breakout session where participants tested the draft rubric with their own use cases and discussed the outcomes of their use case tests. The second and third day consisted of breakout discussions with the following goals:

- Breakout session #2 - Participants were grouped in this breakout session according to how their use cases scored on the rubric in the day one breakout sessions. Four breakout groups were organized to discuss the implications of "preserve most output", "preserve some output", and "preserve little output" when applied to the participant use cases. Participants discussed what components of their simulation workflow should be preserved including the specific model, simulation outputs, and simulation workflow elements.
- Breakout session #3 - Participants were randomly assigned to breakout groups to discuss overarching themes that had repeatedly come up in both virtual workshop #1 and workshop #2. Selected themes included: 1) "How long to preserve simulation output?", 2) "Where are outputs preserved?", 3) "Who pays for the costs of long term data preservation?", and 4) "How do we communicate these challenges with stakeholders, including publishers and funders?"

After the workshop concluded, the project PIs organized common themes and elements found in each use case category, with the goal of developing reference use cases to support rubric score ranges for "Preserve few simulation workflow outputs", "Preserve selected simulation workflow outputs, and "Preserve the majority of simulation workflow outputs". These reference use cases are intended to inform users with guidance on how to proceed, according to the score attained through the rubric. Additionally, the project PIs summarized the common challenges and recommendations that were discussed in breakout session #3, which are intended to inform stakeholders about general concerns related to simulation output preservation. Both of these topics are discussed in more detail below, and links to the general workshop notes are provided near the end of this document.

OUTCOMES

REFERENCE USE CASES ON "WHAT TO PRESERVE"

Three reference use cases on "what to preserve" were compiled as a result of the day one and two breakout discussions. Please find a summary of each use case below.

<u>Knowledge Production</u> - ***Preserve few simulation workflow outputs***

- Summary
    - Doesn't take too many resources to simulate/re-create the runs
    - Generally agreed not to share raw output in this case, keep 2-D diagnostic fields only

    - Sharing software (including model code) has more benefits
- 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
    - Input - initialized data from GFS output, took a sounding and interpolated it to the model grid
    - 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 re-use
    - 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 postprocessing tools where these are available.
- Why should it be preserved
    - 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 non-linear 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

Knowledge Production with use operational numerical weather prediction center products as inputs - ***Preserve selected simulation workflow outputs***

- Summary

    - Who preserves operational simulation output data? The agency that produces the data? The researcher(s) who use them? [point to Broader Issues discussion]
- 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 to preserve?
    - Initiation and assimilation data - Base fields and boundary conditions from HRRRE (HRRR Ensemble). If possible point to the NWP center that produced this data.
    - 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 to preserve?
    - 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.
    - How long to preserve? Where to preserve?
        - Depends on age/relevance
    - The model source code absolutely should be preserved because it's been heavily modified from publicly available versions.

<u>Data Production</u> - *Preserve the majority of simulation workflow outputs*

- Summary
    - Reasons for keeping all/most of the output data
        - Data may be large enough that rerunning simulation or re-doing the post-processing is prohibitively difficult
        - Users expect specific outputs and wouldn't want to rerun the code themselves
        - Model code not shared/shareable (proprietary or research model)
        - People rely on you to run the model to produce outputs because they may not have resources to run the model, or not able to run at full resolution etc.
    - Archive consideration - If you have large data, the access becomes more difficult. You may need special services to support access (e.g.

subsetting). You may also need different approaches to store and access data based on how many users the data will receive.

- Description of use case
    - Using the CMAC model to study ammonia in the atmosphere. Running WRF on CONUS, and perturbe it, so we have 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?
    - Output data preserved with all parameters
    - Notes on how it was produced because it probably won't be possible to reproduce
    - Model code is available, other scripts are small, with some documentation
- Why should it be preserved?
    - 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.

PRELIMINARY CHALLENGES AND RECOMMENDATIONS

Recent years have seen significant development in more specific guidelines on data availability for published research. However, the guidelines are often prohibitive, with requirements suggesting "preserve all the data" which is unclear and unrealistic for a lot of simulation research. The focus of this current EarthCube RCN award is twofold: 1) to develop best practices to assist a researcher in determining what data or software should be preserved in a FAIR aligned repository to communicate knowledge, and 2) communicate that knowledge to publishers to create less prohibitive requirements. These best practices will be in the form of a simulation/experiment descriptor rubric, an accompanying user guide, and a summary report.

Complementary to those specific products and goals, several larger themes have coalesced from discussions at the RCN workshops:

1. There is a crisis in simulation output curation and storage. This primary goal of this project is to develop best practices for deciding what needs to get preserved and communicate those practices clearly to researchers, repositories and publishers. This should decrease the volume

of simulation-related output that needs to be preserved. However, researchers are currently spending a significant portion of their own time dealing with data curation; in some cases, over 50% of their funded time. This is a waste of time and money. Additionally the ecosystem of community repositories to support Atmospheric Science is sparse. We need a coordinated effort to fund personnel to assist researchers in data curation, as well as investment in the needed repository preservation and stewardship services.

2. The primary goal in earth science is replicability, not computational reproducibility. Here, we follow the definitions put forth in this National Academies report: https://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=25303&_ga=2.59905261.1997197855.1557255010-1768681981.1557255010
In other words, the goal is to have both enough information about the workflow and also selected derived outputs to communicate the important environmental characteristics to allow a future researcher to build off of the original study. However, particularly for the large number of highly nonlinear simulation studies, computational reproducibility should not be expected, nor is it needed.

3. The majority of research involving simulations is knowledge production not data production (Baker and Mayernik, 2020*). Most researchers that produce simulation output would love more use of their output products; and many end users** would love more data. But wanting this to be true does not make it so, and 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***. Proposed data production oriented research should include an appropriate budget to support anticipated data preservation and community data access needs.
*https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/ecs2.3191
**The modeling community needs better methods of communicating storage and curation limits to stakeholders and end users. Our hope is the primary products from the current project (e.g., simulation/experiment descriptor rubric) will help facilitate that communication.
***Researchers can continue to locally save as much data as they want.

4. Improved technological capabilities, including cloud storage, doesn't solve all data preservation needs. For example, without data stewardship and curation, cloud storage is nothing more than a modern version of "anonymous FTP". In other words, without investment in data curation personnel, the potential benefits of improved technological capabilities will not be realized.

5. Operational data products which are collected, produced, and disseminated by Numerical Weather Prediction centers, such as the US National Centers for Environmental Prediction, have become essential components in retrospective Atmospheric and Ocean sciences related research. Researchers use these products as initialization and boundary conditions in simulation studies, and are asked to preserve these data to enable replication of their research. Unfortunately many NWP centers don't provide public access to their legacy data archives. This

had led to an inefficient, ad-hoc approach for preserving these data across the research community, with many University research groups collecting and serving various components of these data on locally maintained web servers. In some cases, duplicate copies of data products are stored across many institutions leading to a waste of resources. Additionally, the data volumes of these products tend to be very large (10s to 100s of TBs), and it can be too costly for researchers to preserve these data on local infrastructure, or in trusted repositories as required by many journal publishers. Accordingly, there is a need for NWP centers to preserve and provide open access to the full record of data they collect and produce, to support retrospective research and improve the research to operations relationship. This would take the burden off of the research community to preserve bits and pieces of these products, and would facilitate easier replication of studies that use these products.

LINK TO DRAFT RUBRIC

Version 2.0 rubric - available as PDF or xlsx. The motivating use case for this version of the rubric is: “A rubric to be used to assist a researcher in determining what simulation outputs should be deposited in a FAIR aligned community repository to communicate knowledge.” Version 2.0 was refined according to participant feedback provided in workshop 2. Refinements included: 1) grouping the descriptors into section themes, 2) adding a “big picture” question for each section theme to provide context for what topic the descriptors are targeted at addressing, 3) reordering the descriptors to check if a project falls into the “data production” use case at the beginning of the rubric, 4) adding a “recommended weighting” for each descriptor score, 5) flipping the scores for the“preserve more output” and “preserve less output” descriptor classes, and 6) pointing to reference use cases to inform the rubric users on how to be proceed according the the score they attain. The rubric will continue to be refined according to community feedback moving forward, and a user guide will be developed to better clarify how to use the rubric.

NEXT STEPS

The project PIs plan to engage community stakeholders through targeted meetings and engagements at disciplinary meetings over the next several months. Specifically a town hall to solicit community feedback on the latest version of the rubric will be held at the AMS 2021 annual meeting. From here the project steering committee will determine if and how additional workshops and community engagement would be useful in answering outstanding questions. The PIs are also planning meetings with publishers; representatives from AMS and AGU Publishers have already been involved in the project in various ways (steering committee, workshop presentations and participation). Year 2 of the project will include meetings to discuss improving clarity in journal data requirements and also aligning those requirements with recommendations from this project..

LINKS TO ADDITIONAL MATERIALS

- Project Website
- Workshop Agenda
- Workshop Welcome Video

- General Workshop 2 Notes Document- includes notes from all full workshop sessions, including the plenary discussions and breakout session reports.
- List of workshop registrants
- Plenary Presentations