BAMS Article

What about Model Data?

Best Practices for Preservation and Replicability

Douglas C. Schuster, Matthew S. Mayernik, Gretchen L. Mullendore, and Jared W. Marquis

KEYWORDS:
Climate models; Mesoscale models; Numerical analysis/modeling; Reanalysis data

ABSTRACT: It has become common for researchers to make their data publicly available to meet the data management and accessibility requirements of funding agencies and scientific publishers. However, many researchers face the challenge of determining what data to preserve and share and where to preserve and share those data. This can be especially challenging for those who run dynamical models, which can produce complex, voluminous data outputs, and have not considered what outputs may need to be preserved and shared as part of the project design. This manuscript presents findings from the NSF EarthCube Research Coordination Network project titled "What About Model Data? Best Practices for Preservation and Replicability" (https://modeldatarcn.github.io/). These findings suggest that if the primary goal of sharing data are to communicate knowledge, most simulation-based research projects only need to preserve and share selected model outputs along with the full simulation experiment workflow. One major result of this project has been the development of a rubric, designed to provide guidance for making decisions on what simulation output needs to be preserved and shared in trusted community repositories to achieve the goal of knowledge communication. This rubric, along with use cases for selected projects, provide scientists with guidance on data accessibility requirements in the planning process of research, allowing for more thoughtful development of data management plans and funding requests. Additionally, this rubric can be referred to by publishers for what is expected in terms of data accessibility for publication.

https://doi.org/10.1175/BAMS-D-22-0252.1
Corresponding author: Douglas C. Schuster, schuster@ucar.edu
In final form 24 August 2023

AFFILIATIONS: Schuster and Mayernik—National Center of Atmospheric Research, Boulder, Colorado; **Mullendore**—National Center of Atmospheric Research, Boulder, Colorado, and University of North Dakota, Grand Forks, North Dakota; **Marquis**—University of North Dakota, Grand Forks, North Dakota

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 (Irving 2016). Open and replicable workflows not only strengthen public confidence in the sciences, but also result in more efficient community science (e.g., Alves et al. 2023). Following this push for open science, many publishers, including the American Meteorological Society (AMS) and the American Geophysical Union (AGU), require data availability statements, and many funding agencies expect more effective data management and sharing plans. 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 yet know what to do about data produced as output from computational models. Guidance to researchers varies and is often unclear (Gomes et al. 2022). The simplest solution for replicability would be to "preserve all the data," but simulation data can be prohibitively large, particularly for individual researchers and in a field like atmospheric or oceanic sciences. 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 (Simmonds et al. 2022). Examples of important simulation descriptors include community commitment, simulation workflow accessibility, simulation output accessibility, research feature replicability, and cost of running the simulation workflow compared to the cost of repository data management services.

The primary goal of this article is to share a workflow, involving a rubric and reference use cases, designed to help individual researchers determine what simulation outputs and codes need to be preserved and shared in a trusted community repository for communication of knowledge (e.g., meet data accessibility needs for publishers). Products used to support the workflow were developed through public engagement, including three community workshops funded by the NSF EarthCube Research Coordination Network project titled "What About Model Data? Best Practices for Preservation and Replicability" (MDRCN, https://modeldatarcn.github.io/).

Rubric and use cases

Historically, when individual researchers have developed the project plan for simulation-based research, little thought may have been given to "What and how much data do I need to preserve and share as a result of my project, and in what structure/format to support broad community reuse?" Many researchers simply dumped all model output produced through their projects onto local storage systems or asked repositories to take all of their data "as is" and preserve and share those data indefinitely. As the volumes produced by models have increased, most data repositories no longer accept these types of requests, and local storage accompanied

with either a "contact the author" statement, or a temporary web server to generically serve data are no longer a data reuse solution accepted by a growing number of publishers (Jones et al. 2019) whose data access policies are aligned with the Findability, Accessibility, Interoperability, and Reuse (FAIR) Guiding Principles (Wilkinson et al. 2016). Through workshop discussions it was decided that researchers should complete the following steps (Fig. 1) when a simulation-based project proposal is being formulated: 1) craft a software development and sharing plan that will enable reuse of codes used to drive the full simulation experiment workflow (Mullendore et al. 2021), 2) develop an estimate on what and how much data will need to be preserved and shared as a result of the project by applying the rubric and reference use cases (Schuster et al. 2022), 3) identify and engage a repository to host simulation outputs if the proposal is funded, and 4) include any necessary costs to support long-term data and software preservation and reuse in the proposal budget. By working through these steps early in the research process, the resultant research workflow will ensure that project-generated data will be structured in community accepted formats and rich metadata will be created to enable long-term data discovery and community reuse, software will be structured and documented such that others with domain knowledge can understand and rerun the simulation workflow, and a trusted, community repository (Lin et al. 2020) will be available to preserve and serve the selected data products chosen to communicate knowledge. This avoids the need for researchers or repositories to perform the sometimes costly and time-intensive task of restructuring and reformatting products into community accepted structures and standards after a project is complete and will ensure that researchers will meet publisher and funder expectations for data and software preservation and sharing.

Development of workflow. The ultimate goal of the MDRCN project was to develop guidance and a rubric for authors, funders, and publishers on what data and software elements of simulation-based research need to be preserved and shared to meet community open science requirements and expectations. To achieve this goal, two virtual workshops were held

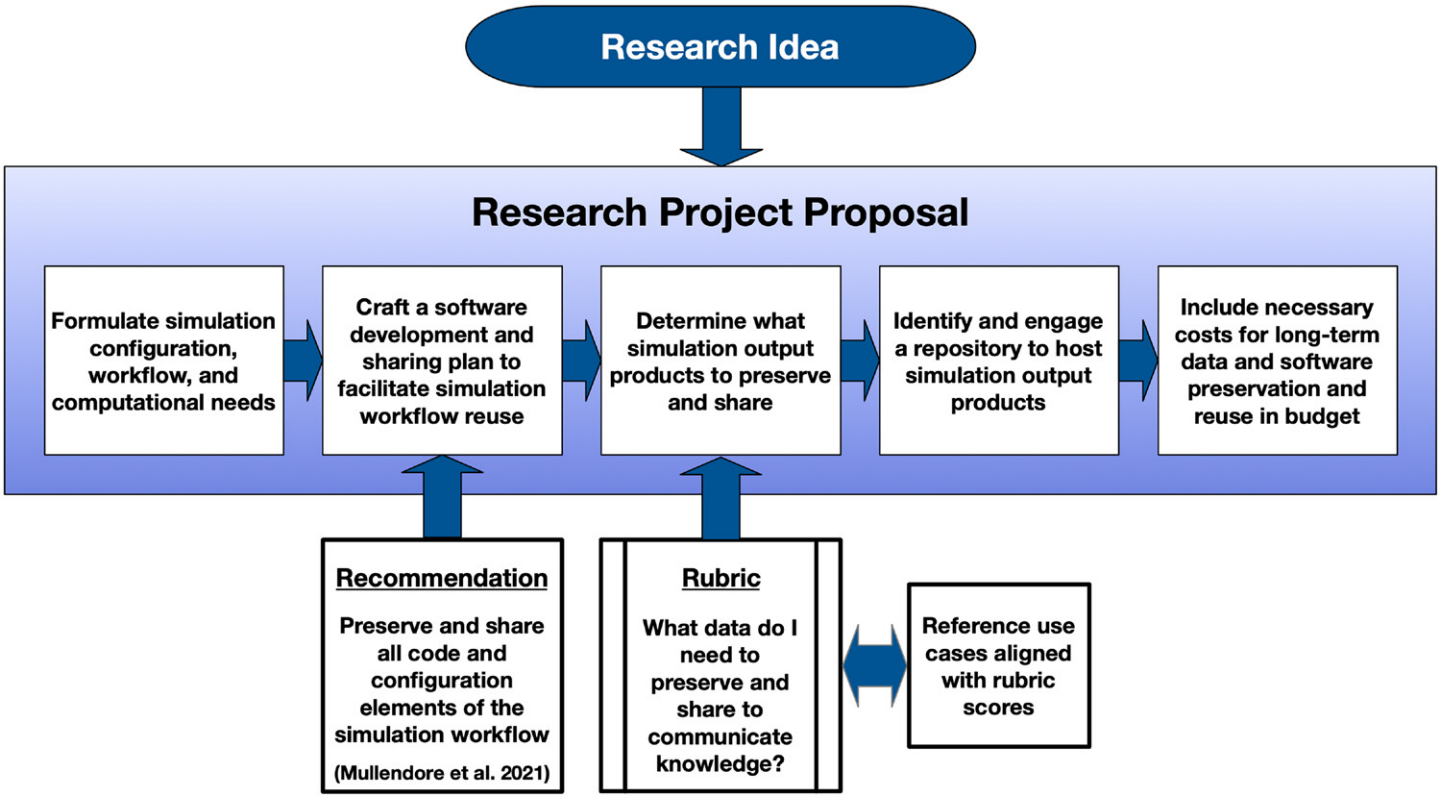

Fig. 1. Data and software management planning during research project proposal development.

in 2020 and one hybrid workshop was held in 2022 with asynchronous product development and review outside of the workshops. Workshops were structured with plenary presentations, followed by multiple breakout session working groups tasked to achieve specific sets of outcomes. Working groups drew upon the knowledge of a diverse array of participants, including simulation-focused researchers from the atmospheric and hydrologic sciences, data curators, librarians, and publishers to co-develop the rubric and associated MDRCN products.

During the first workshop participants worked to craft the rubric by 1) developing lists of model descriptors with definitions; 2) combining the separate descriptor lists into one large list of model descriptors; 3) refining the full descriptor list by combining, distilling, and adding new descriptors; and 4) filling in selected class definitions for each descriptor, with the initial focus on edge cases in order to describe the range of possibilities for a given descriptor. For example, in what cases would one save all of the model output versus in what cases would one save none of the model output for a given descriptor? A list of over 100 descriptors was culled down to 17, and resulted in the first version of the rubric, which was further refined and organized into themes. Although several possible rubric uses were discussed, this 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 trusted community repository to communicate knowledge."

The goal of the second virtual workshop was 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. Breakout groups investigated the implications of "preserve most output," "preserve some output," and "preserve little output" when applied to the participant use cases and discussed what components of their simulation workflow should be preserved including the specific model, simulation outputs, and simulation workflow elements. As a result of this effort, reference use cases were compiled to provide examples on how to proceed according to the score attained through the rubric.

During the third workshop, participants examined issues related to model software and data preservation and sharing that emerged from discussions in the first two workshops. Breakout sessions explored the challenges and possible solutions associated with sustainable curation, determining the lifetime for simulation data, incentivizing data and software sharing, and open and equitable science.

Data production versus knowledge production. We found in the workshops that before discussing the goals and specifics of the rubric and use cases, we first needed to clarify data production versus knowledge production. The majority of research involving simulations is knowledge production, not data production (Baker and Mayernik 2020). 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., Coupled Model Intercomparison Project, CMIP; Eyring et al. 2016), 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 from knowledge production projects than can be sustainably stored in public repositories. Knowledge production research should preserve minimal simulation output in repositories. Further guidance in determining whether a given project is knowledge or data production is provided by following the rubric workflow.

Rubric. The rubric and accompanying use case examples (Schuster et al. 2022) are intended to be used as a tool 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.

The rubric is organized into individual sections categorized by theme and associated big picture question. Each theme is informed by individual descriptor questions to come up with a section total weighted score for that theme. The suggested scoring weights of the rubric descriptors are designed to provide balance between the various rubric section themes and were found to be beneficial during rubric use case testing. Higher scores are associated with "preserve more output" and lower scores are associated with "preserve less output." A complete list of rubric themes, associated big picture questions, descriptor questions by theme and suggested section, and descriptor scoring contributions from the rubric follows. It should be noted that class description text was not developed for many of the "in-between" class ii choice scenarios.

1) **Community Commitment** (Section Total Weighted Score: Min = 3, Max = 18)
   - **Is it anticipated that your simulation workflow outputs will have broad community impact and downstream reuse?**
     (a) Is this simulation output to be used as part of a "highly influential scientific assessment" (HISA) as defined, for example, by White House Office of Management and Budget "Revised Information Quality Bulletin for Peer Review" (April 15 2004)?

     Descriptor Weighted Score by Class (Class #/Score/Class Description)
     (i) **1** - Simulation workflow outputs will not be used in a HISA.
     (ii) **4** - Subset of output may enable fact checking, e.g., all output are not needed, but selected or derived products (e.g., ensemble mean and spread) will provide adequate scientific representation.
     (iii) **6** - Simulation workflow outputs will be used in a HISA. Need to keep output for future fact checking.

     (b) Is this simulation output part of a larger set of experiments that is of value as a whole (e.g., intercomparisons)?
     (i) **1** - Simulation output is not part of a larger set of related experiments.
     (ii) **4** - Subset of data may be more appropriate for some kinds of ensemble experiments.
     (iii) **6** - Simulation output is part of a larger set of related experiments.

     (c) Is this simulation output potentially a community benchmark for comparison?
     (i) **1** - Simulation output is not a benchmark or community reference dataset.
     (ii) **4**
     (iii) **6** - Simulation output is a community reference dataset (e.g., global reanalysis).
2) **Repository Data Accessibility** (Section Total Weighted Score: Min = 2, Max = 12)
   - **Does the trusted community repository that you plan on archiving your data in provide adequate data access capabilities for the volume of data that you plan on depositing?**
     (a) Do bandwidth limitations impede data transfer options from the community data repository expected to archive the simulation output?
     (i) **1** - Data volume is too large to effectively transfer and no data volume reduction capabilities are provided by the repository.
     (ii) **4**
     (iii) **6** - Data volume is small enough, or data volume reduction services are provided by the repository to support effective data transfer.

     (b) Is there a capability to access/use data analysis compute resources collocated with the community data repository where the simulation output will be archived?

(i) **1** - No publicly accessible data analysis compute capabilities are collocated with the data repository expected to host the simulation output.
(ii) **4**
(iii) **6** - Publicly accessible data analysis compute capabilities are collocated with the data repository expected to host the simulation output.

3) **Simulation Workflow Accessibility** (Section Total Weighted Score: Min = 4, Max = 12)
- **Would it be straightforward for others in your academic discipline to rerun your simulation model run workflow steps?**
  (a) How accessible is this particular version of the model/code? Are there intellectual property (IP) barriers, embargo periods for new model development?
    (i) **1** - Community validated version of a highly accessible model was used.
    (ii) **2** - Model source code is shareable, but specific changes were implemented that make it unique. Code is lightly documented.
    (iii) **3** - Model source code is difficult to acquire.
  (b) Is the source code well documented and easy to use?
    (i) **1** - Source code is well documented and easy to install and run.
    (ii) **2**
    (iii) **3** - There is very little supporting documentation. Source code is difficult to understand and manage.
  (c) How specialized of a platform is needed to execute the model (specific hardware, compilers, software libraries needed)?
    (i) **1** - Does not require special hardware, niche software libraries, and licensed compilers to execute. This could include a containerized version of a model.
    (ii) **2**
    (iii) **3** - Requires resources that are more difficult to get access to, e.g., specialized HPC, niche software libraries, and licensed compilers.
  (d) How much effort is it to get and manage all the inputs used by the simulation?
    (i) **1** - Simulation inputs/boundary conditions are easy to acquire and manage.
    (ii) **2**
    (iii) **3** - Simulation inputs/boundary conditions are difficult to acquire and manage and retaining output lowers burden for others who might want to rerun model or use outputs.

4) **Simulation Postprocessing Workflow Accessibility** (Section Total Weighted Score: Min = 3, Max = 9)
- **Would it be straightforward for others in your academic discipline to rerun your simulation postprocessing workflow steps?**
  (a) How accessible is this particular version of the postprocessing code? Are there IP barriers, embargo periods for new model development?
    (i) **1** - Community validated version of a highly accessible postprocessing workflow was used.
    (ii) **2** - Postprocessing source code is shareable, but specific changes were implemented that make it unique. Code is lightly documented.
    (iii) **3** - Postprocessing source code is difficult to acquire.
  (b) Is the postprocessing source code well documented and easy to use?
    (i) **1** - Source code is well documented and easy to install and run.
    (ii) **2**
    (iii) **3** - There is very little supporting documentation. Source code is difficult to understand and manage.
  (c) How specialized of a platform is needed to execute the postprocessing code (specific hardware, compilers, software libraries needed)?

(i) **1** - Does not require special hardware, niche software libraries, and licensed compilers to execute. This could include a containerized version of a post-processing workflow.
(ii) **2**
(iii) **3** - Requires resources that are more difficult to get access to, e.g., specialized HPC, niche software libraries, and licensed compilers.

5) **Research Workflow Output Accessibility** (Section Total Weighted Score: Min = 1, Max = 6)
- **Would it be straightforward for others across academic disciplines to use your simulation workflow outputs?**
  (a) How easy is it to use the outputs outside of the original context? Does it adhere to community standards/conventions (e.g., CF NetCDF)? Are the metadata sufficient for someone else to understand the output?
    (i) **1** - Simulation outputs provided in proprietary format. Obscure or undefined standards make usability and long-term curation difficult.
    (ii) **4**
    (iii) **6** - Simulation outputs structured, formatted, and aligned with community conventions. Data can be easily read by common software and understood in the future.

6) **Research Feature Replicability** (Section Total Weighted Score: Min = 1, Max = 9)
- **Would it be feasible for others in your academic discipline to replicate a physical feature generated through your simulation?**
  (a) Can others replicate specific (atmospheric) features (of given scale) within an acceptable statistical range of error?
    (i) **1** - No issues with specific feature replicability.
    (ii) **6** - Would be difficult to replicate some feature details, but general findings are robust.
    (iii) **9** - Would be difficult to replicate due to nonlinearity of phenomena being studied.

7) **Cost of Running Simulation Workflow** (Section Total Weighted Score: Min = 2, Max = 12)
- **What is the cost to produce your simulation workflow outputs?**
  (a) What is the economic cost (combination of run time and computer access costs) of completing the simulation workflow?
    (i) **1** - Small computational cost and no special platform needs.
    (ii) **4** - Moderate computational cost, but access to needed platforms straightforward.
    (iii) **6** - High computational cost. Need a large compute capability and/or can only be produced with specialized platforms.
  (b) What are the person-hours required to reproduce a simulation dataset?
    (i) **1** - Trivial effort required to replicate simulation for most end users.
    (ii) **4**
    (iii) **6** - Significant time and expertise required to replicate simulation. Likely will require contact with and guidance from original data producer(s).

8) **Repository Data Management Services Cost** (Section Total Weighted Score: Min = 1, Max = 12)
- **What is the cost to archive your output in a trusted community repository to preserve and provide access to your simulation workflow outputs for a minimum period of time?**
  (a) What is the economic cost of curating simulation output in a community repository, for a minimum time period?

(i) **1** - Community repository data curation expenses are prohibitive due to large volume of the expected model outputs.
(ii) **8**
(iii) **12** - Would be inexpensive to curate the complete simulation workflow output for a minimum number of years in a community repository.

To use the rubric, consider a specific simulation workflow during the project formulation phase and select a score according to the class that best fits the characteristics of the simulation workflow for each descriptor found in a section theme. Once scores have been selected for each descriptor, it is recommended that a user total up the score for each section theme to see how each theme contributes to the rubric total score. The rubric total weighted score is intended to inform the user on what to deposit into a repository:

- Rubric Total Weighted Score < 48: **Preserve few simulation workflow outputs**
- 48 ≤ Rubric Total Weighted Score ≤ 72: **Preserve selected simulation workflow outputs**
- Rubric Total Weighted Score > 72: **Preserve the majority of simulation workflow outputs**

As has been illustrated, themes are broken out into individual scoring sections in the rubric, allowing users to view the contributions of each theme to the total rubric score. For example, if one scores high in the "Community Commitment" section of the rubric (e.g., a weighted score of 13–18), this likely indicates the project falls under the "Data Production" concept, where the output produced through a project is intended for reuse by a large number of downstream users, and the user may not need to go through the remainder of the rubric questions. In this case the researcher should plan to preserve and share the majority of the simulation workflow outputs, and the data repository infrastructure to support this should be resourced accordingly to support end user access requirements. Other general themes examined by the rubric include "Accessibility" and "Cost." These are broken out into six separate rubric section themes to investigate whether it would work best and be more cost effective to have end users with domain knowledge examine, understand, and rerun (if needed) the full simulation workflow or reuse simulation workflow outputs to best communicate research knowledge.

The suggested scoring weights of the rubric descriptors are designed to provide balance between the various rubric section themes as can be seen in the "Cost" themes (i.e., sections 7 and 8) where both sections have a maximum total scoring contribution of 12. Certain researchers may have compelling reasons to adjust the weighting for individual section themes based on the goals of a specific project. For example, if it is not reasonable to expect others to replicate a specific physical feature within an acceptable statistical range of error by rerunning the simulation workflow, but the physical feature generated through a simulation is essential to communicating research findings, the researcher may want to increase the weighting of this rubric section (section 6) to nudge the overall score toward the "preserve more output" scoring bins.

Examples of what others have preserved and shared according to their rubric scores are provided in the reference use cases found in Schuster et al. (2022) and described in the following section. These examples are intended to provide a reference in helping others decide what should be preserved and shared for their own project, but it is noted that these each specific decision point is inherently subjective and dependent on project details.

Use cases. As all projects are unique, there is no one solution when it comes to decisions about what data to preserve and share. Projects with identical rubric scores may still decide to preserve different portions of their data output for long-term access. To investigate potential

scenarios for data preservation and sharing, we developed 12 use cases based on discussions with scientists about current and past modeling projects. The first three use cases were developed via discussions with participants in our second project workshop, and the other nine were developed via discussions with modelers based at NCAR. The purpose of the use cases is to give examples of projects in a particular rubric score range.

It is a feature of the use cases that the data collections described therein vary considerably. Some are long lists of a single file type of the same size, while others included multiple file types of varying sizes. They include data collections from general circulation models like the Community Earth System Model (CESM), as well as data from regional and weather models like the Weather Research and Forecast (WRF) Model.

The use cases were based on the following question: How do scientists make decisions about what data and other files to deposit in a repository? Specifically, what data are being preserved and why? The use case template walks through a number of details to get at data preservation decisions. It starts by asking the researcher for a high-level overview of the scientific project, which includes the science goals and basic modeling workflow. Following that background, the template walks through questions about the specific materials that should be preserved and shared in the context of a given project. This includes questions about 1) the data, including the model inputs, raw model output, and processed model output; 2) the software, including asking about the model configuration, preprocessing code, model code, and postprocessing code; and 3) other related information, including documentation and metadata, and any visualizations or image products that are produced by the project and are distinguished from processed output that exists as numerical data. For each of these sections in the use case template, we tried to gather information from the scientists about why they made the retention decisions that they made, along with any specific reasons for these decisions. We also asked scientists about any temporal considerations, such as whether particular products become more or less useful over time.

We asked the use case participants to complete the project's rubric without any assistance from our project team beyond basic instruction. The rubric scores for these use case examples generally were consistent with their choices about what to deposit in the repository. This indicated that our rubric is aligned with community expectations for data preservation and sharing, and that the use cases could serve as helpful examples for rubric scores within a similar range. The use cases range from "preserve few simulation outputs," such as idealized process studies in which the goal is knowledge production and there is more value in sharing model configurations and codes than data, to "preserve the majority of simulation outputs," which is applicable to model intercomparison projects where data are being generated specifically to enable reuse by others.

Most use cases in our current collection, however, fall into the middle category, "preserve selected simulation outputs." This middle category is the most challenging in terms of the decisions to be made about data preservation. In these cases, the modeling cannot be replicated via a trivial rerun of the model, either due to complexity of the model, large data volumes, specialized hardware needed, or a combination of these and other factors. But what does "selected simulation outputs" mean? Our use cases were developed to contextualize the rubric by showing strategies that modelers are using to reduce overall data volume, while still providing enough useful data to enable transparency, replication, and follow-on studies.

These approaches include, but are not limited to the following:

- *Compression*: Various techniques exist to reduce volume of data files via lossless or lossy compression (Duben et al. 2019). Compression is widely used, including both general-purpose compression formats, such as zip, and compression tools built for specific file formats, such as NetCDF. Compression is useful to reduce overall data volume

and should be used to complement the other strategies described below when feasible. On its own, however, it does not solve questions about what data to preserve in a repository.

- *Lower resolution*: Reducing resolution is another common approach for simulation projects. This may include reducing spatial resolution, such as running a model and analyzing the outputs using a 3 km grid but preserving data using 9 km grid. In the case of temporal resolution, it is common to run models using time steps of 1 h or shorter but then generate daily or monthly averages for data presentation and preservation.
- *Selecting specific variables*: Many analyses of weather and climate simulations focus on a small subset of the raw model output and/or derived variables. In a number of our use cases, modelers archived only derived variables that were central to the research project and did not archive the raw model fields and/or excluded raw model fields not important for their analyses, e.g., surface model variables.
- *Excluding miscellaneous files*: Also generally excluded from data archives are other files that are necessary to run and evaluate model output, but were not the focus of analysis. This includes restart files and log files that may be generated to monitor the performance of a model but are not relevant once the model has completed running.
- *Preserving only the files used to create figures*: The data used to generate the figures in scientific papers are generally a highly processed and carefully analyzed distillation of the raw model output, and often have gone through one or more of the approaches noted in the previous bullets. As such, it is a common approach to only archive these final processed files used to generate the figures and tables in scientific articles, rather than the raw model output.

A few other high-level takeaways emerged from the use cases. First, the main purpose for depositing data for most scientists who participated in the discussions was to fulfill publisher requirements for data preservation and sharing. This speaks to the importance of publisher requirements to advance open data goals, but also the need for publishers to provide guidance for researchers in how to meet these data sharing requirements. Other purposes that were mentioned for depositing data into a repository include to provide data for some specific user community and to reduce the work required by others to generate and use particular model outputs. This last motivation was noted specifically for complex models where the outputs are very large in volume.

We recognize that the use cases would benefit from additional questions that interrogate the broader impacts of the research and data. Specifically, do the curation decisions consider the scientific needs of and effects on historically oppressed communities, given the historic passive and active exclusion of these communities from most scientific efforts? We were not able to bring to maturity an effort to add these questions to the use case template, but we strongly recommend such considerations are included in curation decision-making to produce the best outcomes.

Summary and conclusions

As funding agencies, publishers, and research institutions push for more open and effective science, many researchers question “What about model data?” when making decisions about data preservation and sharing. While several community efforts are underway to incentivize and enable open science, including the FAIR Principles on making data Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016), the CARE Principles for Indigenous Data Governance (Carroll et al. 2020), and the TRUST Principles which provide guidance for repositories (Lin et al. 2020), these efforts do not provide specific guidelines on simulation data. Recent projects such as AtMoDat (Atmospheric Model Data; Ganske et al. 2021, 2022)

and the Coupled Model Intercomparison Projects (CMIP; Petrie et al. 2021) have provided standards for structuring model output and metadata, yet little guidance on what model outputs to preserve and share existed.

This article presents the outcomes from a community effort to answer the question "What about model data?" when making decisions about data preservation and sharing. Through a series of workshops, town halls, and other community engagements, a rubric and use cases were developed to help researchers determine which simulation outputs and codes should be preserved and shared in a trusted community repository for communication of knowledge (e.g., when publishing an article). We recommend that researchers use the rubric at the proposal stage of projects. The rubric (and use cases) will help the researchers estimate future archiving needs for both project outputs and software and can help make workflow decisions that will facilitate easier curation later on. We also recommend engaging with a repository at this stage to estimate costs and include these costs in the proposal. This sort of planning is already encouraged as part of the Data Sharing Plans required by most major funding agencies.

When this project commenced, there was no clear guidance on what should be done with model data; many organizations were defaulting to "save everything." Much progress has been made since, both as part of this effort as well as within the broader community. Recently AGU (https://data.agu.org/resources/agu-data-software-sharing-guidance) and AMS (https://www.ametsoc.org/index.cfm/ams/publications/ethical-guidelines-and-ams-policies/data-and-software-policy-guidelines-for-ams-publications/) have both updated their data and software sharing policies and started referring journal articles authors to use the MDRCN products described in this article for guidance on what model outputs should be preserved and shared to support open science expectations. There do remain unsolved issues (e.g., who should pay for storage and curation?) and iteration on best practices should continue. Additionally, the onus of data and software management currently falls on the researchers themselves, who often do not have the curation expertise or the time to make their data and software understandable (Mullendore et al. 2021). However, we have taken a first step. The products and approaches detailed here should serve as an important foundation for the continuing collective effort toward achieving open and replicable science.

Acknowledgments. The authors thank the many workshop participants who contributed significantly to this effort. We would also like to the thank the project steering committee members; in particular, thanks to Clark Evans and Adam Clark for help on use case development, and thanks to Elisa Murillo for broader impacts discussions and language. We also thank David Eby for contributing to the development of the use cases. This project was funded by the NSF EarthCube program, NSF Awards 1929757 and 1929773. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977.

Data availability statement. All products created through this project are openly available from NCAR's Geoscience Data Exchange at https://doi.org/10.5065/g936-q118 as referenced in Schuster et al. (2022).

References

Alves, J.-H., H. Tolman, A. Roland, A. Abdolali, F. Ardhuin, G. Mann, A. Chawla, and J. Smith, 2023: NOAA's Great Lakes Wave Prediction System: A successful framework for accelerating the transition of innovations to operations. *Bull. Amer. Meteor. Soc.*, **104**, E837–E850, https://doi.org/10.1175/BAMS-D-22-0094.1.

Baker, K. S., and M. S. Mayernik, 2020: Disentangling knowledge production and data production. *Ecosphere*, **11**, e03191, https://doi.org/10.1002/ecs2.3191.

Carroll, S., and Coauthors, 2020: The CARE principles for indigenous data governance. *Data Sci. J.*, **19**, 43, https://doi.org/10.5334/dsj-2020-043.

Düben, P. D., M. Leutbecher, and P. Bauer, 2019: New methods for data storage of model output from ensemble simulations. *Mon. Wea. Rev.*, **147**, 677–689, https://doi.org/10.1175/MWR-D-18-0170.1.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Ganske, A., and Coauthors, 2021: Atmodat standard (v3.0). World Data Center for Climate (WDCC) at DKRZ, 40 pp., https://doi.org/10.35095/WDCC/atmodat_standard_en_v3_0.

——, A. Heil, A. Lammert, J. Kretzschmar, and J. Quaas, 2022: Publication of atmospheric model data using the ATMODAT standard. *Meteor. Z.*, **31**, 493–504, https://doi.org/10.1127/metz/2022/1118.

Gomes, D. G. E., and Coauthors, 2022: Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc. Biol. Sci.*, **289**, 20221113, https://doi.org/10.1098/rspb.2022.1113.

Irving, D., 2016: A minimum standard for publishing computational results in the weather and climate sciences. *Bull. Amer. Meteor. Soc.*, **97**, 1149–1158, https://doi.org/10.1175/BAMS-D-15-00010.1.

Jones, L., R. Grant, and I. Hrynaszkiewicz, 2019: Implementing publisher policies that inform, support and encourage authors to share data: Two case studies. *Insights UKSG J.*, **32**, 11, https://doi.org/10.1629/uksg.463.

Lin, D., and Coauthors, 2020: The TRUST principles for digital repositories. *Sci. Data*, **7**, 144, https://doi.org/10.1038/s41597-020-0486-7.

Mullendore, G. L., M. S. Mayernik, and D. C. Schuster, 2021: Open science expectations for simulation-based research. *Front. Climate*, **3**, 763420, https://doi.org/10.3389/fclim.2021.763420.

Petrie, R., and Coauthors, 2021: Coordinating an operational data distribution network for CMIP6 data. *Geosci. Model Dev.*, **14**, 629–644, https://doi.org/10.5194/gmd-14-629-2021.

Schuster, D. C., M. S. Mayernik, and G. L. Mullendore, 2022: Products developed through the "What about Model Data? Determining Best Practices for Preservation and Replicability, Earthcube Research Coordination Network" project. UCAR/NCAR GDEX, https://gdex.ucar.edu/dataset/id/6962fde0-9f65-4530-9320-76c42866c821.html.

Simmonds, M. B., and Coauthors, 2022: Guidelines for publicly archiving terrestrial model data to enhance usability, intercomparison, and synthesis. *Data Sci. J.*, **21**, 3, https://doi.org/10.5334/dsj-2022-003.

Wilkinson, M. D., and Coauthors, 2016: The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018, https://doi.org/10.1038/sdata.2016.18.