# Guidelines for Publicly Archiving Terrestrial Model Data to Enhance Usability, Intercomparison, and Synthesis

RESEARCH PAPER


MAEGEN B. SIMMONDS
WILLIAM J. RILEY
DEBORAH A. AGARWAL
XINGYUAN CHEN
SHREYAS CHOLIA
ROBERT CRYSTAL-ORNELAS
ETHAN T. COON
DIPANKAR DWIVEDI
VALERIE C. HENDRIX
MAOYI HUANG
AHMAD JAN
ZARINE KAKALIA
JITENDRA KUMAR
CHARLES D. KOVEN
LI LI
MARIO MELARA
LAVANYA RAMAKRISHNAN
DANIEL M. RICCIUTO
ANTHONY P. WALKER
WEI ZHI
QING ZHU
CHARULEKA VARADHARAJAN

*Author affiliations can be found in the back matter of this article

CORRESPONDING AUTHOR:
**Charuleka Varadharajan**
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA
*cvaradharajan@lbl.gov*



## ABSTRACT

Scientific communities are increasingly publishing data to evaluate, accredit, and build on published research. However, guidelines for curating data for publication are sparse for model-related research, limiting the usability of archived simulation data. In particular, there are no established guidelines for archiving data related to terrestrial models that simulate land processes and their coupled interactions with climate. Terrestrial modelers have a unique set of challenges when publishing data due to the diversity of scientific domains, research questions, and the types and scales of simulations. Researchers in the U.S. Department of Energy's (DOE) projects use a variety of multiscale models to advance robust predictions of terrestrial and subsurface ecosystem processes. Here, we synthesize archiving needs for data associated with different DOE models, and provide guidelines for publishing terrestrial model data components following FAIR (Findable, Accessible, Interoperable, Reusable) principles. The guidelines recommend archiving model inputs and testing data used in final simulation runs along with associated codes, workflow scripts, and metadata in public repositories. Researchers should consider archiving model outputs if they are within the storage limits of the repository. We also provide considerations for how to bundle files into different data publications with citable digital object identifiers. Finally, we identify repository features and tools that would enable storage and reuse of model data. Given the diversity of DOE terrestrial models, these guidelines are transferable to other model types and will enable efficient reuse of simulation data for purposes such as model intercomparisons, initialization, benchmarking, synthesis, and comparisons with field observations.



KEYWORDS:
data management; archiving guidelines; FAIR; terrestrial models; simulations

TO CITE THIS ARTICLE:
Simmonds, MB, Riley, WJ, Agarwal, DA, Chen, X, Cholia, S, Crystal-Ornelas, R, Coon, ET, Dwivedi, D, Hendrix, VC, Huang, M, Jan, A, Kakalia, Z, Kumar, J, Koven, CD, Li, L, Melara, M, Ramakrishnan, L, Ricciuto, DM, Walker, AP, Zhi, W, Zhu, Q and Varadharajan, C. 2022. Guidelines for Publicly Archiving Terrestrial Model Data to Enhance Usability, Intercomparison, and Synthesis. *Data Science Journal*, 21: 3, pp. 1–18. DOI: *https://doi.org/10.5334/dsj-2022-003*

## 1. INTRODUCTION

Data management and stewardship in scientific research are critical to accelerating knowledge discovery across domains. Recently, the scientific community has promoted the use of Findable, Accessible, Interoperable, and Reproducible (FAIR) principles to make data from research activities broadly available (Wilkinson *et al.*, 2016; Stall *et al.*, 2019). The FAIR principles outline how to make data and information easy to "discover, access, interoperate, and sensibly re-use, with proper citation" (Wilkinson *et al.*, 2016). This can be achieved in part by archiving data supporting the results of scientific research in public repositories for long-term preservation and discoverability. In addition, adopting data or metadata standards and reporting formats that specify preferred file formats and variable names will improve reusability (Crystal-Ornelas *et al.*, 2021). In most cases, community engagement and consensus is requisite to adopting these standards and guidelines, and building cohesiveness among archived datasets (Sansone *et al.*, 2019).

Many current standards are targeted towards observational and experimental datasets (*https://fairsharing.org/standards/*). In contrast, guidelines on archival of model data are limited, but have proven to be extremely useful when available (e.g, Jones *et al.*, 2016; Fer *et al.*, 2021). For example, researchers involved in the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project (CMIP) activities built consensus on requirements for their archived data to be traceable, reproducible, and usable for scientific purposes (Jones *et al.*, 2016; Durack *et al.*, 2018). The global climate model data are available for broader use outside of the CMIP network via the distributed data archive, Earth System Grid Federation (ESGF) (*https://esgf.llnl.gov/*). The scientific objectives of each CMIP project informed the design of the data archives and the standardization of the datasets providing for example, detailed documentation of experimental conditions, requested variables, data reference syntax and controlled vocabulary, general structure and format of the data, and file directory system organization (*https://pcmdi.llnl.gov/CMIP6/Guide/dataUsers.html*). These archives enabled much of the model-based research in the Intergovernmental Panel on Climate Change assessments, such as improving estimates of the carbon cycle (Arora *et al.*, 2020).

Terrestrial models (alternatively known as land models) are a broad class of Earth science numerical models that simulate land dynamics and fluxes of energy, water, carbon, and nutrients (Fisher and Koven, 2020). Terrestrial models can be coupled with global Earth system models and other regional-scale models, or run 'offline' at site, watershed, river basin, continental, or global scales (Sood and Smakhtin, 2015). Terrestrial modeling datasets lack guidelines for public archiving, and have a unique set of attributes that make building consensus on standardized archiving protocols challenging. First, the data are very diverse since they are used to address a broad range of questions across different scientific domains spanning climate, hydrology, biogeochemistry, and ecology. Moreover, these models can be used at vastly different spatial and temporal scales to study ecosystem processes. For example these models can be used to investigate the drivers of the terrestrial carbon sink at global scales (Riley, Zhu and Tang, 2018), as well as to understand the fate of riverine chemistry at local to watershed scales (Dwivedi *et al.*, 2018; Jan, Coon and Painter, 2021). Finally, model data can have many components, including output files of various dimensions and resolutions (e.g., final raw outputs, spin-up output files, restart files, test data files, and higher level outputs corresponding to figures); a variety of metadata files (embedded within output files such as those in NetCDF formats or external to the data files); visualization files; model code; input files (e.g., model parameters, climate forcing data, surface data); scripts for model set-up and initialization; code to calculate and assign input parameters; post-processing; and visualizations.

The terrestrial modeling community would benefit from a set of guidelines for curating model data for long-term archival. However, there is no current community consensus on answers to several important questions related to publishing model data, including 1) what model-related data are worth archiving, 2) how much storage space is needed and what are suitable repositories to host such data, and 3) what are best practices for curating the datasets and associated files (e.g., model code and pre- and post-processing scripts). Guidelines for curating modeling data for long-term public archival would enable their reuse for purposes such as spinning up new simulations, model synthesis and intercomparisons, comparisons of model predictions with observational data, and informing experimental designs that reduce prediction uncertainty.

Terrestrial models are used in U.S. Department of Energy (DOE) research to advance a robust, predictive understanding of climate impacts on ecosystem processes such as carbon cycle changes caused by warming (Huang *et al.*, 2019; Riley *et al.*, 2021), vegetation dynamics (Mekonnen *et al.*, 2019), or watershed responses to disturbances such as early snowmelt and droughts (Hubbard *et al.*, 2018). The Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) is a data repository established to serve as the long-term steward of environmental research data sponsored by the DOE (Varadharajan *et al.*, 2019). ESS-DIVE stores heterogeneous data types (e.g., hydrological, biogeochemical, ecological, climate, remote sensing) generated by observational, experimental, and modeling activities and seeks to enable data discovery and reuse by partnering with the science community. Several terrestrial model datasets from DOE research are publicly available on ESS-DIVE (e.g., Arora et al., 2019; Dwivedi, 2019; Fung, 1993; Hilton and Baker, 2018; Walker, De Kauwe, et al., 2018) and other archives such as the ESGF. In this study, the ESS-DIVE team worked collaboratively with a diverse set of modeling researchers across the DOE community to determine guidelines for long-term archival of terrestrial model data in public repositories.

The main objectives of this study were to (1) synthesize current practices and recommendations across the Earth Science modeling and data repository communities for archiving model data, (2) assess requirements for public archiving, synthesis, and utilization of a diverse selection of terrestrial model data, and (3) provide pragmatic recommendations about best practices for curating scientifically useful model datasets, including those associated with scientific publications, towards enabling reproducibility of modeling workflows and data reuse for purposes such as model results intercomparison and synthesis. Below we describe our review of previous approaches to storing model data and our recommendations on archiving terrestrial model data. Although the study was designed to inform the ESS-DIVE repository policies, the guidelines are broadly applicable to other model types and data archives given the diversity of terrestrial model data considered in this study. To our knowledge, this is the first study that provides recommendations for archiving different components of model data for scientific purposes. Such guidelines are necessary as publication of model datasets is expected to grow significantly as journals and funding sources expand their requirements, and needs special consideration due to the volume and complexity of data associated with typical simulations.

## 2. METHODS

### 2.1 REVIEW OF EXISTING MODEL DATA ARCHIVING GUIDELINES

First, we researched capabilities of existing data systems that support Earth science model or large data archiving including the ESGF, the National Aeronautics and Space Administration (NASA) Earth Observing System Data and Information System (EOSDIS) Distributed Active Archive Centers (DAACs) (*https://earthdata.nasa.gov/eosdis/daacs*), the National Center for Atmospheric Research (NCAR) Research Data Archive (RDA) (*https://rda.ucar.edu/*), the Earth Observatory Lab (EOL) data archive (*https://data.eol.ucar.edu/*), and the National Science Foundation (NSF) Arctic Data Center (*https://arcticdata.io/*). We also reviewed general-purpose repositories Dryad (*https://datadryad.org/*), Zenodo (*https://zenodo.org/*) and ESS-DIVE that accept large data files. The review considered current storage capacities and guidelines provided by these systems for contributing model-specific and other types of data.

Additionally, we reviewed existing guidelines for archiving model data from the National Science Foundation (NSF) EarthCube Model Data Research Coordination Network (RCN) and the American Geophysical Union (AGU). The NSF EarthCube model data RCN (EarthCube-RCN) group has been researching and hosting workshops on best practices for geoscientific model data preservation and reproducibility (*https://modeldatarcn.github.io/*) and developed a rubric as a decision-support tool for researchers choosing how much of their simulation workflow output (raw outputs to post-processed outputs) to publicly archive in a FAIR-aligned data repository. We also reviewed journal-specific guidance on publishing modeling data on the AGU website (*https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-for-Authors*) (Hanson, 2020). The results from the review are summarized in section 3.1.

### 2.2 UNDERSTANDING COMMUNITY TERRESTRIAL MODEL DATA ARCHIVING NEEDS AND PERSPECTIVES

We determined needs for archiving, sharing, and utilizing archived data across a broad range of terrestrial models used in DOE research projects. For this study, we gathered input from 12 researchers who work across multiple DOE projects and institutions and use a diverse set of modeling codes to address a wide variety of science questions. We collected input using a form with a set of questions to determine 1) what types of models are currently used in DOE research? 2) what are approximate data volumes and file types generated for different simulations, 3) what components of model data are considered scientifically useful to archive? 4) how long does archived model data remain useful for the scientific community? 5) how do modelers currently archive their data? 6) what features should data repositories support to enable storage and reuse of archived model data in the future? (see supplementary information for the full list of questions). The questions regarding the value of archiving different model data components and importance of different repository features used a rank measure on a five-point scale ranging from 1 (not important) to 5 (highly important).

We also conducted discussions with researchers who had published model data in five scientific publications to determine their workflows and priorities for archiving data (Zhu, Riley and Tang, 2017; Walker *et al*., 2019; Zhi *et al*., 2019; Jan, Coon and Painter, 2020; Koven *et al*., 2020).

### 2.3. DETERMINING MODEL DATA ARCHIVAL GUIDELINES

We aggregated the input provided by the modelers by taking average scores for questions that had an importance rank, and by tabulating responses for the other questions. We additionally considered input from the follow-on discussions and reviewed the data and code availability statements in the 5 journal publications (Supplemental Table 1) to determine the range of data archiving practices across modelers and to determine practical challenges associated with publishing simulation data. We then drafted an initial version of the guidelines based on our review of other Earth science model archiving practices (Section 3.1) and the input from the modelers participating in this study. The guidelines were finalized with community consensus on what was practical to archive given potential uses of the data and capabilities of current repositories that accept model data.

## 3. RESULTS

### 3.1 SYNTHESIS OF MODEL DATA ARCHIVING CAPABILITIES AND GUIDELINES ACROSS DATA CENTERS AND ORGANIZATIONS IN THE EARTH SCIENCE COMMUNITY

***Table 1*** summarizes properties of seven data centers used by the Earth science community that we reviewed in terms of their data publication storage limitations and the availability of guidelines for curating a model data publication or other archiving best practices. At the time this study was conducted, only the NSF Arctic Data Center (ADC) and NASA's Oak Ridge National Laboratory Distributed Active Archive Center (ORNL-DAAC) for Biogeochemical Dynamics provided some guidance that could be used by data contributors to publish model-related data, code, or scripts.

The ADC provides guidelines on metadata associated with software (includes models); files to include for models and scripts; file organization and formats; and considerations for archiving large datasets including model output data (*https://arcticdata.io/submit/*). The ORNL-DAAC provides guidelines for submission of model code or scripts and recommend including model code, documentation specifying the model name and version, model process representation and, as appropriate, a description of model lineage, sample input and output (*https://daac.ornl.gov/submit/*). Their guidelines specify acceptable file formats, including common model output and input file formats (e.g., NetCDF, HDF5, GeoTIFF, shapefile, CSV), and suggests including files necessary "to represent a complete, and reproducible, body of work". They also provide general guidelines on data and file organization; file-level metadata; file formats and naming; types of files expected in a data publication such as data files, supplemental files (including photos, reports, or metadata), documentation, code (if applicable), and the published paper or manuscript draft (if applicable; *https://daac.ornl.gov/datamanagement/#best_practices*). The ADC and ORNL-DAAC do not explicitly describe which components of model data files should be archived, such as model inputs, testing data, outputs, model code, and scripts.

| | | PROVIDES DATA CONTRIBUTOR GUIDELINES | |
|---|---|---|---|
| **DATA CENTER** | **STORAGE LIMIT PER DATA PUBLICATION** | **MODEL-DATA SPECIFIC?** | **OTHER?** |
| National Science Foundation Arctic Data Center | No limit | Yes | Yes |
| Oak Ridge National Laboratory DAAC | NA[1] | Yes | Yes |
| NASA's Earth Observing System Data and Information System (EOSDIS) | NA[1] | NA[1] | Yes |
| U.S. DOE ESS-DIVE | 10GB/500 GB[2] | No | Yes |
| Dryad | 300 GB[2] | No | Yes |
| Zenodo | 50 GB | No | No |
| Earth System Grid Federation (ESGF) | NA[1] | NA[1] | NA[1] |

**Table 1** Summary of data centers and their data publication storage limitations, and resources for data contributors on best practices for curating data packages, modeling related and in general.
[1] NA: Not available, i.e. no public information found.
[2] Limit on size of individual files. For ESS-DIVE, 10GB is the default file size limit, and can be increased upto 500GB by request. Files >500GB are considered upon review.

Of the other data systems, the EOSDIS has standards and templates, specifies file formats (netCDF/HDF5), and provides a curation service for data publication based on the user's service level. Dryad (*https://datadryad.org/stash/best_practices*) and ESS-DIVE (*https://docs.ess-dive.lbl.gov/contributing-data/data-submission-guidelines*) have guidelines for dataset-level metadata and submissions. ESS-DIVE also has formats for specific data types and we note that the guidelines presented here will be adopted for its model datasets in the future.

The NSF EarthCube rubric allows modelers to respond to a series of questions that assess potential uses of simulation data ranging from data production to knowledge production (Baker and Mayernik, 2020). A score is calculated based on their responses indicating how much of the outputs (all data to minimal data) should be archived. The level of importance of eight themes in the simulation workflow is considered in the rubric: (1) data production for downstream uses (e.g., CMIP would score highly); (2) repository data accessibility; (3) simulation workflow accessibility (e.g., system requirements, code availability and ease of use); (4) post-processing workflow accessibility (e.g., system requirements, ease of use of scripts and documentation); (5) simulation data accessibility (e.g., follows community standards, ease of use with metadata and documentation); (6) research feature reproducibility; (7) cost of running simulations; and (8) cost of data repository storage and management services.

The AGU guidelines only require that the data that supports the research and visualizations presented in a journal article submission be archived in a FAIR-aligned data repository. They provide tiered options (acceptable, good, best) for citing and describing the model, configuration, and parameters within the journal article, and what to do regarding data corresponding to tables and figures, and model data output.

## 3.2 DIVERSITY IN TERRESTRIAL MODELING DATA

Several terrestrial models are used in DOE research projects for standalone or coupled simulations (***Table 2***), but the majority of the codes used are sponsored by the DOE. The DOE models are run at different spatial (soil pore to global) and temporal scales and resolutions (***Table 3***). Each simulation can contain 5 to a few million files with average file sizes ranging from 100 MB to 2 TB (mean = 280 GB/file, median = 3 GB/file), and currently require hundreds of megabytes to a few hundred terabytes of storage space (mean = 28 TB/modeler, median = 650 GB/modeler). While most modelers used HDF5 or netCDF file formats to save model outputs and metadata, some also used other common formats such as text, comma separated value, or DAT files as well as formats unique to certain models (e.g. Tecplot files, XML, MESH, VTK, PY, EXO). There are numerous types of scripts used in a modeling workflow, ranging from single analyses for specific papers to scripts used every time for preparing model inputs. These scripts similarly can be in a diversity of file formats including those produced by workflow tools such as Jupyter Notebooks (*http://jupyter.org*).

**Table 2** Summary of the standalone terrestrial models used by 12 researchers participating in this study. Coupled models (e.g., ELM-FATES and ELM-PLOTRAN) are not listed but were also considered in evaluating archiving needs.

| MODEL ACRONYM | MODEL NAME (ORGANIZATION) | REFERENCES | DESCRIPTION |
|---|---|---|---|
| ELM | Energy Exascale Earth System Model (E3SM) Land Model (DOE) | Golaz *et al.* (2019); *https://e3sm.org/* | Land model component of the E3SM Earth System Model |
| FATES | Functionally Assembled Terrestrial Ecosystem Simulator (DOE) | Koven *et al.* (2020); *https://github.com/NGEET/fates-release* | Size and age-structured vegetation demographic model within a land surface model and can be coupled with an Earth system model |
| PFLOTRAN | Parallel Flow and Transport (DOE) | Hammond, Lichtner and Mills (2014); *https://www.pflotran.org* | Parallel reactive flow and transport model for subsurface hydrobiogeochemical processes |
| ATS | Advanced Terrestrial Simulator (DOE) | Coon *et al.* (2020); *https://amanzi.github.io/ats/* | An integrated, distributed watershed hydrology model including surface and subsurface flow, energy transport, reactive transport, and ecohydrology. |
| CrunchFlow | N/A (DOE) | Steefel and Molins (2009) | Model for simulating multicomponent multi-dimensional reactive transport in porous media |
| MAAT | Multi-Assumption Architecture & Testbed (DOE) | Walker, Ye, *et al.* (2018); *https://github.com/walkeranthonyp/MAAT* | Modular terrestrial ecosystem process modeling framework for building multiple models that vary in process representation/ hypotheses. |
| CLM | Community Land Model (NCAR) | Lawrence *et al.* (2019); *https://www.cesm.ucar.edu/models/clm/* | Land model for the Community Earth System Model (CESM), a fully-coupled global climate model |
| ED2 | Ecosystem Demography Biosphere Model (NSF/NASA) | Longo *et al.*, (2019); *https://github.com/EDmodel/ED2* | Size- and age- structured terrestrial biosphere model |
| PRMS | Precipitation Runoff Modeling System (USGS) | Markstrom *et al.* (2015); *https://www.usgs.gov/software/precipitation-runoff-modeling-system-prms* | Deterministic process-based model developed to evaluate the impacts of climate and land use on streamflow and watershed hydrology. |
| SWAT | Soil and Water Assessment Tool (USDA/Texas A&M University) | Bieger *et al.* (2017); *https://swat.tamu.edu/* | Watershed to river basin-scale model used to simulate the quality and quantity of surface and ground water and predict the environmental impact of land use, land management practices, and climate change. |
| LPJ-GUESS | Lund-Potsdam-Jena General Ecosystem Simulator (Lund University) | Smith, Prentice and Sykes (2001); *https://web.nateko.lu.se/lpj-guess/* | Dynamic vegetation-terrestrial ecosystem model for regional or global studies |
| GDAY | Generic Decomposition and Yield | Comins and McMurtrie (1993); *https://github.com/mdekauwe/GDAY* | Stand-scale ecosystem model that simulates carbon, nitrogen, and water dynamics. |
| SDGVM | Sheffield Dynamic Global Vegetation Model (Sheffield University) | Woodward and Lomas (2004); *https://bitbucket.org/walkeranthonyp/sdgvm/* | Terrestrial biosphere carbon cycle model for ecosystem to global scale simulations. Simple size and age structure. |
| OpenFOAM | N/A (OpenFOAM foundation) | *https://openfoam.org/* | Computational fluid dynamics open source software |
| CALAND | California Natural and Working Lands Carbon and Greenhouse Gas Model (California Natural Resources Agency) | Di Vittorio and Simmonds (2019); *https://doi.org/10.5281/zenodo.3256727*. | Carbon stock and flux model that simulates the effects of various management practices, land use and land cover change, wildfire, and climate change on ecosystem carbon dynamics across all California lands |

### 3.3 PERSPECTIVES ON BEST PRACTICES FOR PRESERVATION AND REUSE

There was broad consensus amongst the modelers participating in this study that model input files, metadata, and scripts used in the workflow or analysis should be archived for the data to be usable and traceable (***Figure 1a***). Many of the modelers considered it useful, as defined by an importance rank of 3 or higher (somewhat important to very important), to

**Table 3** Estimates of archiving needs for typical spatial and temporal representations of simulation data from DOE terrestrial models, which are the most commonly-used models by the researchers in this study. Note that the same models are often run at different spatial extents (e.g., site to global) and temporal duration (e.g., weeks to centuries).
[1] Note that *"ensembles"* of simulations were not considered in this survey, except in the total annual storage needs reported.
[2] This could represent either the simulation temporal resolution, or output file temporal resolution.
[3] Here we use Land Surface Model *"LSMs"* to include both standard CMIP-style Earth System Models (e.g. ELM) and more complex vegetation phenology models (e.g. FATES).
[4] Note that *"point"* is used to indicate a single vertical column of cells or otherwise a single location in horizontal space.

| DETAILS FOR TYPICAL SIMULATION[1] TO BE ARCHIVED | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **MODEL** | **SPATIAL RESOLUTION OR REPRESENTATION** | **SPATIAL EXTENT** | **TEMPORAL RESOLUTION[2]** | **TEMPORAL DURATION** | **NO. OF FILES** | **MEAN FILE SIZE (GB)** | **TYPES OF FILE FORMATS** | **TOTAL ANNUAL STORAGE NEEDS (GB)** |
| Multiple LSMs[3] | Point[4] | point | daily | 200 yrs | 300 | 0.1 | CSV | 50 |
| ELM | point | point | hourly, daily | 10 – 20 yrs | 20 | 0.004 | netCDF | 3 |
| ELM | 1/2° – 2° | global | monthly | 250 yrs | 2500 | 0.2 | netCDF | 15000 |
| ELM-FATES | point, ~1 km, ~1 degree | point, regional, and global modes | sub-daily, monthly | ~500 yrs | 1K – 10K | 50 | netCDF | 1000 |
| FATES | point | point | <hourly | 10 yrs | 70 | 3 | netCDF | 2000 |
| ELM-PFLOTRAN | 1 – 100 m | 100 m – 10 km | hourly/daily | 10+ yrs | 10 – 100 | 10 | HDF5, netCDF | 1000 |
| PFLOTRAN | <1 m | 5-6 km | <hourly | 30 yrs | 5 | 1000 | HDF5 | 10000 |
| ATS | 100 m – 250 m | 10 km | daily | 10 – 100 yrs | 20 | 100 | XML + HDF5, CSV | 1000 |
| ATS | <1 – 100 m | 10 m – 10 km | daily | 10 – 100 yrs | | 2 | XML + HDF5 | 1000 |
| ATS | 0.25 m | 25 m | daily | 100 yrs | 50 – 200 | | XML + HDF5 | 10 |
| CrunchFlow | <1 m | <1 km | <hourly | 30 days | 100 | 0.001 | TXT | 1 |

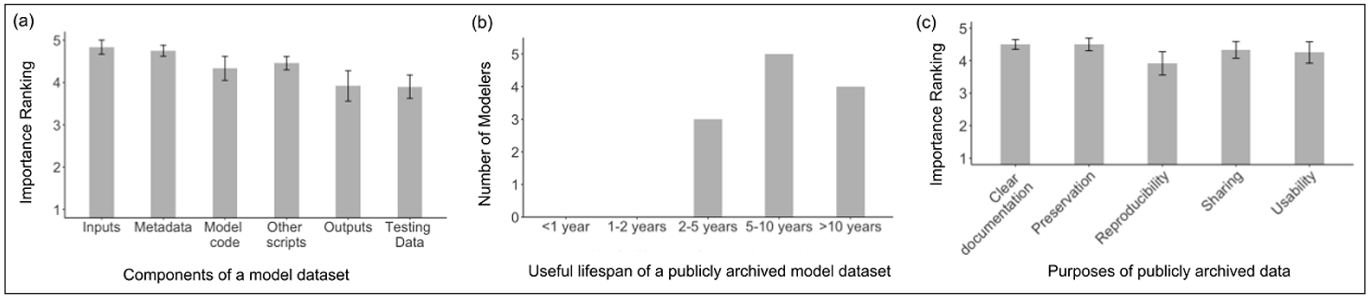


**Figure 1** Perspectives from a group of 12 U.S. Department of Energy terrestrial model researchers of (**a**) archiving different components of model data in a public repository (**b**) the period of time over which publicly archived model data remain useful, and (**c**) purposes served by archiving model data in a public repository. The importance ranking for (a) and (c) are shown as 1 (not important at all) to 5 (extremely important), and represent average importance scores across 12 researchers.

archive the entire workflow including model code (10 out of 12 modelers; mean importance = 4.3), the outputs corresponding to final simulations (8 out of 12; mean importance = 3.9), model input parameters and forcings (11 out of 12; mean importance = 4.8), and scripts for pre-processing and post-processing, model configuration, and analysis (11 out of 12; mean importance = 4.5).

However, there were diverse opinions on the specifics of which model data files are worth preserving. If possible, modelers preferred to archive the majority of model data from final simulation runs (e.g., raw and aggregated outputs), with the exception of files already stored in a repository or public codebase separately with preexisting digital object identifiers (DOIs) or files produced from intermediate steps that are easily reproduced. However, modelers sometimes preferred to only archive high-level outputs corresponding to results presented in a journal article, because the full set of model outputs may be too large to store in most data repositories and can be reproduced with affordable computational cost. Fewer modelers ranked archiving of testing data as important (6 out of 12 modelers; mean importance = 3.9). The rationale provided was that frequently the validation datasets used to test model performance are archived elsewhere and can be referenced in the metadata of a published dataset.

Aside from the simulation files used to derive the published figures and tables in a journal article, modelers also run spin-up simulations and in some cases a small number of higher-resolution simulations than the final simulations used for publication. There was consensus that spin-up simulations are not a high priority for archiving, but that it is worthwhile to publicly archive restart files that allow a model data user to rerun a segment of a simulation in the event that they want to reanalyze the data.

Besides the data files, most modelers (11 out of 12) preferred that specific scripts used for analysis should be archived. However, if a modeler anticipates running analogous simulations many times, then the scripts and model outputs can be archived separately with DOIs, allowing the outputs to be updated over time. Ten out of 12 modelers agreed that model code should be publicly archived for various reasons, but they had different perspectives about where and for how long it should be archived given that model codes can evolve significantly over time. One consideration for storing model code in a data repository was the need for long-term preservation with citable DOIs. Alternatively, most models are currently stored in collaborative software development and sharing platforms (e.g., GitHub, Bitbucket) that interface with Version Control Systems (VCS). Although VCS platforms were considered to be useful for versioning and interaction on model development, releases, tracking issues and bug-fixes, there was concern that VCS systems are not guaranteed to be long-term archives. An approach that some modelers used to balance the needs for long-term preservation and practical software development was to archive tagged releases of model codes with a DOI by utilizing an established partnership between the GitHub software platform and Zenodo data archive.

The modelers also had different perspectives on how long publicly archived model data would remain useful (***Figure 1b***), spanning short (2-5 years; 3 out of 12 modelers), medium (5-10 years; 5 out of 12) and long (>10 years; 4 out of 12) time periods. However, they generally agreed that it was important (as indicated by an importance rank of 3 or higher) to archive data in public repositories for many purposes (***Figure 1c***) that includes sharing (11 out of 12 modelers; mean importance = 4.3), preservation (11 out of 12; mean importance = 4.5), clear documentation of the model runs (12 out of 12; mean importance = 4.5); ensuring reproducibility of workflows (8 out of 12; mean importance = 3.9), and reuse of model data (7 out of 12; mean importance = 4.2). Ultimately, all the modelers agreed that standards for archiving model data are needed to ensure its usability and were willing to learn new organizational guidelines or standardized reporting formats for model data.

### 3.4 CURRENT MODELER PRACTICES FOR PUBLIC ARCHIVING OF TERRESTRIAL MODEL DATA RELATED TO JOURNAL ARTICLES

Model data archived in a FAIR-aligned data repository for the 5 journal articles considered in this study include metadata and model outputs (4 out of 5 articles), followed by model inputs, testing data, model code, and a user guide or readme files (3 out of 5 articles for each component). Three of the 5 journal articles published model-related files under a single DOI, while 2 articles archived multiple datasets. Two of the researchers archived scripts and Jupyter Notebooks for generating inputs, post-processing model output, generating figures, or initiating model simulations. One researcher archived file-level metadata that defined variables and file-naming conventions for machine-readability. Three out of 5 authors made the model code available using GitHub (with or without a Commit ID referenced in article) or Zenodo. Most researchers referenced the storage location of the model data and code in the Data and Code availability section(s) of the paper (Supplemental Table 1).

### 3.5 RECOMMENDED GUIDELINES FOR TERRESTRIAL MODEL DATA ARCHIVING ASSOCIATED WITH SCIENTIFIC PUBLICATION

We recommend the following guidelines for organizing model-related files for simulation workflows. Components of archived model data should include metadata, data files, and optionally user guides, which are described in further detail below. We also provide a decision tree to determine whether to group components into one data publication or split into multiple datasets (***Figure 2***). This decision tree helps address the challenges associated with choosing how much model data to save and other considerations in publishing model-related files, such as varying authorship for different model data components and repository storage limitations.

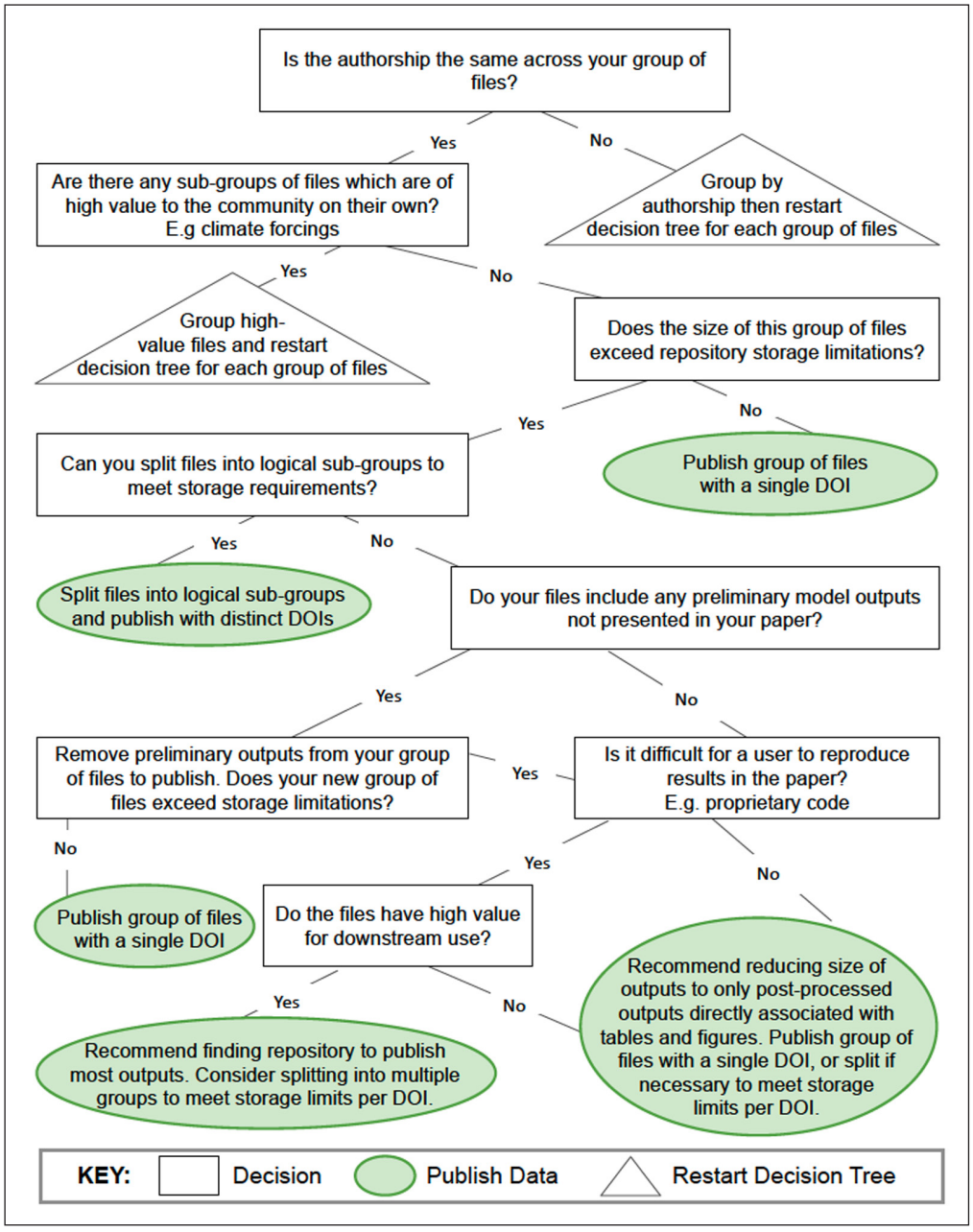


**Figure 2** Decision tree for determining recommended approach for grouping model-related files for public archiving.

1. **Metadata** – This refers to pertinent information about data and code archived (e.g., abstract, geographical and temporal extents), as well as description of the files being archived with links to other DOI-issued publications within the entire simulation workflow, as applicable.
2. **Required Data Files** – Archived datasets should specify or include model inputs, outputs, code, and scripts depending on whether the data are published elsewhere or exceed repository dataset size limits. File names should be unique and can use an intuitive file naming nomenclature to help with discoverability. File names should only contain letters, numbers, hyphens, and underscores, should not contain spaces, and should not rely on case-sensitive file systems.
    a. **Model Inputs** – Input files should be included unless publicly available elsewhere, in which case a hyperlink to the specific input files (e.g., climate forcings, meshes, soil parameterizations) should be provided in the metadata and user guide. Use open-sourced formats such as comma separated value (.csv) or NetCDF (.nc) formats where possible.
    b. **Model Outputs** – Archive all model outputs if the size of the data files are within the repository storage limitations. This output should include the raw and post-processed data, and if associated with a scientific publication the data that support the main findings, tables, and figures. If the size of the model output exceeds repository storage limitations, evaluate recommendations based on the decision tree (***Figure 2***) on which data to publish. Use open-sourced formats such as comma separated value (.csv) or NetCDF (.nc) formats where possible.

   c. **Model Code** – Include source code(s) used to generate results in paper unless the code is publicly available elsewhere (e.g., GitHub or Zenodo), in which case include specific version, hash information, or citation allowing the exact source code to be recovered. Include links to any external model codes in the metadata and user guide. If published on GitHub, provide the commit hash associated with the specific version. If available, include a reference (with DOI) to the tagged release in an established data repository.
   d. **Scripts** – Include run scripts if they are necessary for running the model to generate published results. Optionally also include scripts necessary for reproducing the parameters and model configuration for the simulations and input files, for post-processing model outputs to produce the results (e.g., tables and figures in a publication), and for executing the entire workflow used to generate the model results.
3. **Optional Files** –
   a. ***File-level metadata (FLMD)** – Include descriptions of all the data files as one file catalog (e.g., Velliquette *et al.*, 2021). Optionally also include one data dictionary for each file type within the data publication describing columns and variables.
   b. **Model Testing Data** – Include data files of observations from each location simulated to produce the results in the paper in an open source format (e.g., CSV). If the data are publicly available in another repository, include a reference (with DOI) in the metadata and user guide.
   c. **Documentation or user guide** – Include a readme file (e.g., pdf) for each site-specific or large-scale simulation and provide details on the model name and version number, and required data or code dependencies. Also include a citation for the model code and licensing information if applicable.
4. **Use in publications** – If publishing model results, cite and include links to the data and code publication(s) in the Data or Code Availability section. Include the citations of the dataset and code publication(s) with DOI(s) in the references section. Examples of data or code Availability statements associated with the journal articles researched in this study are provided in Supplemental Table 1.

Further details on these guidelines are described on the ESS-DIVE Community Space on GitHub (*https://github.com/ess-dive-community/essdive-model-data-archiving-guidelines*). The GitHub site also allows for users of these guidelines to provide feedback, and for tracking any future revisions to the guidelines (Crystal-Ornelas *et al.*, 2021).

## 4. DISCUSSION

### 4.1 PUBLIC ARCHIVAL OF MODEL DATA USING RECOMMENDED GUIDELINES

The guidelines we propose are a first step towards improving search capabilities and discovery within model data files, and support the following scientific purposes: 1) repeat the simulations with the same models for traceability and evaluation of the main findings (e.g., data in figures and tables of scientific publication); 2) evaluate published model simulations against observations and other models to gain understanding about model discrepancies and evaluate model uncertainty; and 3) leverage the work for model intercomparison; synthesis of results for meta-analysis or model ensembles; developing new simulations, (e.g., with new spatial domains or input parameters); and for training. We also provide a decision flowchart as a framework for choosing how much of the model data workflow to archive, particularly when storage limitation is an issue or when flexibility is required for supporting a variety of model data archival options.

The guidelines can enable reproducibility of complex scientific workflows that include data ingestion to generate parameter files or other model inputs, running a model multiple times, and analysis of model outputs. We note that these guidelines are specifically focused on establishing provenance of the data used in simulations and enabling reproducibility of modeling workflows, which is sometimes referred to as traceability (Digiampietri et al., 2007). The guidelines are not sufficient to ensure computational (bitwise) reproducibility of model results, which is challenging because of the complexity of modeling codes and diversity of compute architectures and software libraries (National Academies of Sciences, Engineering,

and Medicine, 2019; Goeva, Stoudt and Trisovic, 2020). The ambiguity in how modelers perceive reproducibility may have been a reason for why it received a lower importance rank compared to other purposes for archiving model data (***Figure 1c***).

Although the guidelines were developed in partnership with DOE scientists, the breadth of models used in their research make our recommendations broadly applicable to archival of data from other mechanistic process-based models. In comparison to pre-existing model data guidelines (EarthCube-RCN, NSF Arctic Data Center, ORNL-DAAC), our recommendations strike a balance between the complexity of considerations needed to properly archive the various components of model data and a need for the guidelines to be practical and useful for scientists. We have created additional user-friendly documentation using the GitBooks feature of GitHub (Crystal-Ornelas *et al.*, 2021) to enable adoption of these guidelines (*https://ess-dive.gitbook.io/model-data-archiving-guidelines/*).

## 4.2 ENABLING MODEL DATA INTERCOMPARISON, WORKFLOW REPRODUCIBILITY AND SYNTHESIS

Archiving model data using such guidelines can facilitate coordinated Model Intercomparison Projects (MIPs) and synthesis of data from individual simulation experiments. Data standardization is necessary for MIP efforts since the primary goal is to compare model outputs. Standards have been established and developed for Earth system model outputs, including standardized variable names, units, and other metadata, as part of intercomparison efforts such as CMIP and the Distributed Model Inter-comparison Project (DMIP) (Smith *et al.*, 2013). However, terrestrial models have typically not conformed to standards in their direct outputs. Sometimes a translation tool, such as the Climate Model Output Rewriter (CMOR; *https://pcmdi.github.io/cmor-site/*), can be used to translate the native model output to a standards-compliant format. Most of these toolchains are designed around large-scale modeling exercises and may not be applicable to small-scale studies, such as individual manuscripts or even niche intercomparison efforts. For example the permafrost model intercomparison effort was a small MIP effort undertaken as part of the permafrost carbon network (McGuire *et al.*, 2018), which produced a large number of manuscripts but with only a subset of the models having a standardized output. Another small MIP example that succeeded in establishing an internally-consistent standardized format is the Free Air $CO_2$ Enrichment Model-Data Synthesis (FACE-MDS), and in this instance, the format took several months to develop (Walker *et al.*, 2014; Walker, Kauwe, *et al.*, 2018; Walker, Yang, *et al.*, 2018). Furthermore, conflicting standards exist between MIPs with similar objectives such as the North American Carbon Program (NACP) Multi-scale synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Huntzinger *et al.*, 2013) and the Global Carbon Project (GCP; Friedlingstein *et al.*, 2020), which complicates efforts to converge towards a standard. The guidelines presented here are the first steps toward resolving these issues and enabling model intercomparisons. Further work is needed to develop more complex terrestrial model data standards for variable conventions, units, and other aspects relevant to specific MIP efforts.

Archiving model data from individual studies can also enable reproducibility of their workflows and reuse or synthesis of the data for other analyses. Individual researchers may pre-process data or parameterize and calibrate models in different ways, but the use of computational tools such as Jupyter Notebooks allows the archiving of such analyses and runtime scripts in a more transparent way for subsequent researchers to build on. For example, Koven *et al.* (2020) synthesized multiple datasets on plant traits alongside other model drivers such as site-observed meteorology to run multiple instances of the FATES vegetation model and analyze its outputs. The workflow was captured in Jupyter-based scripts that were cited and archived in Zenodo with a DOI (Koven, 2020). We note that despite the integration between Zenodo and GitHub, many projects hosted on GitHub do not take the extra step to archive their content into long-term data repositories (Crystal-Ornelas *et al.*, 2021), and we highlight this as a key step toward long-term model data reuse and accreditation.

The use cases provided by the modelers participating in this study highlight some of the valuable outcomes of using a common methodology for curating terrestrial model data for publication (e.g., standardized output formats, variable names and units), thereby enabling synthesis of modeled and observational datasets. Coordination in the approaches used for curating the model data would also support the development of products for coupled models (Phillips *et al.*, 2017).

### 4.3 DESIRED DATA REPOSITORY FEATURES AND CYBERINFRASTRUCTURE TOOLS FOR IMPROVING MODEL DATA STORAGE AND REUSE

There are several cyberinfrastructure and data management challenges related to archiving model data. First, data are rapidly increasing in volume and complexity. For example, there is increasing use of ensemble model runs (e.g., Harp *et al.*, 2016; Koven *et al.*, 2020; Cromwell *et al.*, 2021) and very high-resolution simulations (Bisht *et al.*, 2017; Zhu *et al.*, 2020, 2021), which are critical for watershed models and the global land-surface modeling community and result in very large output data volumes. Second, the data are extremely diverse across scientific domains and spatial and temporal scales. Third, there is a disconnect between model and observational data, and fragmentation between workflows attempting to integrate these data. This problem is difficult for many modeling workflows that require manual retrieval of data from multiple sources and subsequent pre-processing for use in modeling analyses.

The immediate need for many researchers to publicly archive data associated with scientific publications to meet journal and funding requirements. Archiving big data on cloud platforms with public accessibility to analytical tools is becoming a trend and is especially important for models with terabyte to petabyte scale outputs. However, cloud storage can be prohibitively expensive and can incur recurring costs for storage, egress and access. Unfortunately, many data repositories are not designed to meet the expected annual storage needs of current large simulations (order of ~1-10 TB; ***Table 2***) and need additional capabilities for enabling archival of datasets at this scale. First, a significant expansion of repository storage capacities is needed to support individual dataset sizes of hundreds of gigabytes to terabytes. In addition, improvements in data transfer capabilities, such as the use of programmatic web services or file transfer services (e.g. Globus; *https://www.globus.org/*) are needed to support large data ingestion and download. Data replication is needed for redundancy and long-term preservation, but poses a challenge with larger datasets. Data repositories also need to support versioning of the numerous files generated from model simulations over the course of a project, especially since many modelers change their archived data several times during manuscript preparation to final publication and beyond.

A long-term need for modelers is to have a more seamless process for publication, such as a *model-to-archive pipeline* that would constitute various data repository resources and services that can support consistent archiving of diverse model data. For example, a support tool for assisting modelers in following the recommended guidelines would be useful, such as an interface or scripts that automate the writing and organization of the files comprising the simulation workflow components. This tool could be model-specific and assemble all the required data for publication in specified formats by extracting subsets of model simulations corresponding to specific runs, locations, variables, or figures. Such a tool could also be extended to enable more advanced querying, utilization, and synthesis of model datasets beyond the metadata. Another example is a tool that provides support for containerized images (e.g. Docker; *https://www.docker.com/*) containing model codes and associated data, which can make it a lot easier to reproduce model data and results.

In an effort to improve the transparency of model-data integration and data provenance, repositories should consider mechanisms to provide links to internal and external datasets that are part of the pre- or post-processing workflows. For example, interoperability is needed between the repository or data center storing a researcher's model data, and other systems that store data needed for generating model inputs or testing datasets. Linking datasets across repositories require consensus on which existing metadata standards to use and how to identify the different relationships and linked data types needed to provide a comprehensive view of the model dataset. A longer-term need is a *data-to-model pipeline* that can enable integration of observational data available across data systems with simulation codes, which would dramatically improve the efficiency of modeling workflows. Such a pipeline could focus on supporting data formats that are typically used in model simulations (e.g., NetCDF and HDF5), including the ability to retrieve data through programmatic means and export data into these formats.

There is a pressing need for more repositories to archive the growing volumes of model datasets and design solutions to address the challenges posed by their size, diversity and interoperability requirements. Community engagement with modelers is essential to identify archival priorities and to develop practically feasible guidelines for curating standardized model data publications that follow data management best practices.

## 5. CONCLUSIONS

The terrestrial modeling community needs to publish standardized simulation datasets in repositories that can support large data archival, model data reuse, and integration with other data centers. In this study, we synthesize archiving needs across several terrestrial models used by U.S. DOE researchers and propose an initial set of guidelines that specify how different model data components (e.g., model inputs, outputs, scripts, metadata) should be archived. The guidelines serve different scientific purposes, including traceability of published research and reuse of data for model intercomparisons and synthesis efforts. We also provide guidance for splitting model data into multiple datasets depending on repository capabilities, authorship, and other considerations. Finally, we identify short-term and long-term repository features and software tools to assist modelers with archiving and sharing simulation data and codes, and improving their scientific workflows. These guidelines are broadly applicable beyond the models considered in this study, and are urgently needed given increasing volumes of published terrestrial model data.

## DATA ACCESSIBILITY STATEMENT

The data presented in this publication including the recommended guidelines are published in the ESS-DIVE repository (Simmonds *et al*., 2021). Future updates to the guidelines will be managed and available through the ESS-DIVE community GitHub repository (*https://github.com/ess-dive-community/essdive-model-data-archiving-guidelines*).

## ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplemental Information.** File containing the list of questions and details of journal articles used in this study to assess model data archiving needs. DOI: *https://doi.org/10.5334/dsj-2022-003.s1*

## ACKNOWLEDGEMENTS

ESS-DIVE is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Earth and Environmental Sciences Division, Data Management program under contract number DE-AC02-05CH11231. ESS-DIVE uses resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. ORNL is managed by UT-Battelle, LLC, for the DOE under contract DE-AC05-1008 00OR22725. We thank two anonymous reviewers whose feedback helped improve this manuscript, and William Collins (LBNL) for his thoughtful insights on model data archiving.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

Conceptualization: MS, WR, CV

Data curation: MS

Validation: WR, CK, AW, SC, AJ, QZ, WZ, LL, DD, JK, XC, SP, DR, EC, MH, KM, CV

Formal Analysis: MS, CV

Visualization: ZK, MS, CV

Writing – original draft: MS, CV, WR

Writing – review and editing: MS, CV, WR, CK, AW, RCO, MM, SC, AJ, QZ, WZ, LL, DD, JK, XC, SP, DR, EC, MH, KM, DA

Funding Acquisition: DA, CV

## AUTHOR AFFILIATIONS

**Maegen B. Simmonds** *orcid.org/0000-0001-7796-7154*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA; Now at: Pivot Bio, 2910 Seventh Street, Berkeley, CA 94710, USA

**William J. Riley** *orcid.org/0000-0002-4615-2304*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Deborah A. Agarwal** *orcid.org/0000-0001-5045-2396*
Computing Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Xingyuan Chen** *orcid.org/0000-0003-1928-5555*
Pacific Northwest National Laboratory, Richland, WA, USA

**Shreyas Cholia** *orcid.org/0000-0002-4775-8201*
Computing Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Robert Crystal-Ornelas** *orcid.org/0000-0002-6339-1139*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Ethan T. Coon** *orcid.org/0000-0001-8124-9622*
Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

**Dipankar Dwivedi** *orcid.org/0000-0003-1788-1900*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Valerie C. Hendrix** *orcid.org/0000-0001-9061-8952*
Computing Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Maoyi Huang** *orcid.org/0000-0001-9154-9485*
Pacific Northwest National Laboratory, Richland, WA, USA; Now at: Weather Program Office, Oceanic and Atmospheric Research, National Oceanic and Atmospheric Administration, Silver Spring, MD 20910

**Ahmad Jan** *orcid.org/0000-0003-2781-7857*
Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA; Now at: National Oceanic and Atmospheric Administration Affiliate, Office of Water Prediction, National Water Center, Tuscaloosa AL 35401 (United States)

**Zarine Kakalia** *orcid.org/0000-0001-9045-1160*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA; College of Natural Resources, University of California Berkeley, Berkeley, CA 94720, USA

**Jitendra Kumar** *orcid.org/0000-0002-0159-0546*
Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

**Charles D. Koven** *orcid.org/0000-0002-3367-0065*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Li Li** *orcid.org/0000-0002-1641-3710*
Department of Civil and Environmental Engineering, The Pennsylvania State University, State College, PA 16802, USA

**Mario Melara** *orcid.org/0000-0002-2343-5280*
Computing Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Lavanya Ramakrishnan** *orcid.org/0000-0003-1761-4132*
Computing Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Daniel M. Ricciuto** *orcid.org/0000-0002-3668-3021*
Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

**Anthony P. Walker** *orcid.org/0000-0003-0557-5594*
Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

**Wei Zhi** *orcid.org/0000-0001-5485-1095*
Department of Civil and Environmental Engineering, The Pennsylvania State University, State College, PA 16802, USA

**Qing Zhu** *orcid.org/0000-0003-2441-944X*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

**Charuleka Varadharajan** *orcid.org/0000-0002-4142-3224*
Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

## REFERENCES

**Arora, B, Bill, M, Conrad, M, Dong, W, Faybishenko, B, Molins, S, Spycher, N, Steefel, C, Tokunaga, T, Wan, J** and **Williams, K.** 2019. Influence of hydrological, biogeochemical and temperature transients on subsurface carbon fluxes in a flood plain environment, Biogeochemistry: Dataset. *Watershed Function SFA, ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.21952/WTR/1506937* on 2021-11-15.

**Arora, VK,** et al. 2020. 'Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models'. *Biogeosciences*, 17(16): 4173–4222. DOI: *https://doi.org/10.5194/bg-17-4173-2020*

**Baker, KS** and **Mayernik, MS.** 2020. 'Disentangling knowledge production and data production'. *Ecosphere*, 11(7): e03191. DOI: *https://doi.org/10.1002/ecs2.3191*

**Bieger, K,** et al. 2017. 'Introduction to SWAT+, A Completely Restructured Version of the Soil and Water Assessment Tool'. *JAWRA Journal of the American Water Resources Association*, 53(1): 115–130. DOI: *https://doi.org/10.1111/1752-1688.12482*

**Bisht, G,** et al. 2017. 'Coupling a three-dimensional subsurface flow and transport model with a land surface model to simulate stream–aquifer–land interactions (CP v1.0)'. *Geoscientific Model Development*, 10(12): 4539–4562. DOI: *https://doi.org/10.5194/gmd-10-4539-2017*

**Comins, HN** and **McMurtrie, RE.** 1993. 'Long‐Term Response of Nutrient‐Limited Forests to CO"2 Enrichment; Equilibrium Behavior of Plant‐Soil Models'. *Ecological Applications*, 3(4): 666–681. DOI: *https://doi.org/10.2307/1942099*

**Coon, ET,** et al. 2020. 'Coupling surface flow and subsurface flow in complex soil structures using mimetic finite differences'. *Advances in Water Resources*, 144: 103701. DOI: *https://doi.org/10.1016/j.advwatres.2020.103701*

**Cromwell, E,** et al. 2021. 'Estimating Watershed Subsurface Permeability From Stream Discharge Data Using Deep Neural Networks'. *Frontiers in Earth Science*, 9: 3. DOI: *https://doi.org/10.3389/feart.2021.613011*

**Crystal-Ornelas, R,** et al. 2021. 'A Guide to Using GitHub for Developing and Versioning Data Standards and Reporting Formats'. *Earth and Space Science*, 8(8): e2021EA001797. DOI: *https://doi.org/10.1029/2021EA001797*

**Digiampietri, L, Medeiros, C, Setubal, J** and **Barga, R.** 2007. Traceability Mechanisms for Bioinformatics Scientific Workflows. *AAAI Workshop – Technical Report*.

**Durack, P,** et al. 2018. 'Toward Standardized Data Sets for Climate Model Experimentation'. *Eos*, 2 July. Available at: *http://eos.org/science-updates/toward-standardized-data-sets-for-climate-model-experimentation* (Accessed: 15 November 2021). DOI: *https://doi.org/10.1029/2018EO101751*

**Dwivedi, D.** 2019. Hot spots and hot moments of nitrogen in a riparian corridor, Water Resources Research: Dataset. *Watershed Function SFA, ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.21952/WTR/1506939* on 2021-11-15. DOI: *https://doi.org/10.21952/WTR/1506939*

**Dwivedi, D,** et al. 2018. 'Hot Spots and Hot Moments of Nitrogen in a Riparian Corridor'. *Water Resources Research*, 54(1): 205–222. DOI: *https://doi.org/10.1002/2017WR022346*

**Fer, I,** et al. 2021. 'Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration'. *Global Change Biology*, 27(1): 13–26. DOI: *https://doi.org/10.1111/gcb.15409*

**Fisher, RA** and **Koven, CD.** 2020. 'Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems'. *Journal of Advances in Modeling Earth Systems*, 12(4): e2018MS001453. DOI: *https://doi.org/10.1029/2018MS001453*

**Friedlingstein, P,** et al. 2020. 'Global Carbon Budget 2020'. *Earth System Science Data*, 12(4): 3269–3340. DOI: *https://doi.org/10.5194/essd-12-3269-2020*

**Fung, I.** 1993. Goddard Institute for Space Studies (GISS) 3-Dimensional (3-D) Global Tracer Transport Model (DB1006). Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (USA) Goddard Institute for Space Studies (GISS), NASA. *ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.3334/CDIAC/CYC.DB1006* on 2021-11-15. DOI: *https://doi.org/10.3334/CDIAC/cyc.db1006*

**Goeva, A, Stoudt, S** and **Trisovic, A.** 2020. 'Toward Reproducible and Extensible Research: from Values to Action'. *Harvard Data Science Review*, 2(4). DOI: *https://doi.org/10.1162/99608f92.1cc3d72a*

**Golaz, J-C,** et al. 2019. 'The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution'. *Journal of Advances in Modeling Earth Systems*, 11(7): 2089–2129. DOI: *https://doi.org/10.1029/2018MS001603*

**Hammond, GE, Lichtner, PC** and **Mills, RT.** 2014. 'Evaluating the performance of parallel subsurface simulators: An illustrative example with PFLOTRAN'. *Water Resour. Res.*, 50(1): 208–228. DOI: *https://doi.org/10.1002/2012WR013483*

**Hanson, B.** 2020. "Data policies and practices for AGU publications for models and model output". *Presented at the National Science Foundation EarthCube Model Data RCN Workshop*.

**Harp, DR,** et al. 2016. 'Effect of soil property uncertainties on permafrost thaw projections: a calibration-constrained analysis'. *The Cryosphere*, 10(1): 341–358. DOI: *https://doi.org/10.5194/tc-10-341-2016*

**Hilton, TW** and **Baker, IT.** 2018. SiB3 simulations of gross primary productivity(GPP) and carbonyl sulfide (COS) plant flux. *Scaling from Flux Towers to Ecosystem Models: Regional Constraints on Carbon Cycle Processes from Atmospheric Carbonyl Sulfide, ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1460838* on 2021-11-15. DOI: *https://doi.org/10.15485/1460838*

**Huang, Y,** et al. 2019. 'Realized ecological forecast through an interactive Ecological Platform for Assimilating Data (EcoPAD, v1.0) into models'. *Geoscientific Model Development*, 12(3): 1119–1137. DOI: *https://doi.org/10.5194/gmd-12-1119-2019*

**Hubbard, SS,** et al. 2018. 'The East River, Colorado, Watershed: A Mountainous Community Testbed for Improving Predictive Understanding of Multiscale Hydrological–Biogeochemical Dynamics'. *Vadose Zone Journal*, 17. DOI: *https://doi.org/10.2136/vzj2018.03.0061*

**Huntzinger, DN,** et al. 2013. 'The North American Carbon Program Multi-Scale Synthesis and Terrestrial Model Intercomparison Project – Part 1: Overview and experimental design'. *Geoscientific Model Development*, 6(6): 2121–2133. DOI: *https://doi.org/10.5194/gmd-6-2121-2013*

**Jan, A, Coon, ET** and **Painter, SL.** 2020. 'Evaluating integrated surface/subsurface permafrost thermal hydrology models in ATS (v0.88) against observations from a polygonal tundra site'. *Geoscientific Model Development*, 13(5): 2259–2276. DOI: *https://doi.org/10.5194/gmd-13-2259-2020*

**Jan, A, Coon, ET** and **Painter, SL.** 2021. 'Toward more mechanistic representations of biogeochemical processes in river networks: Implementation and demonstration of a multiscale model'. *Environmental Modelling & Software*, 145: 105166. DOI: *https://doi.org/10.1016/j.envsoft.2021.105166*

**Jones, CD,** et al. 2016. 'C4MIP – The Coupled Climate–Carbon Cycle Model Intercomparison Project: experimental protocol for CMIP6'. *Geoscientific Model Development*, 9(8): 2853–2880. DOI: *https://doi.org/10.5194/gmd-9-2853-2016*

**Koven, C.** 2020. ckoven/runscripts: version 1.0 of ckoven/runscripts. *Zenodo*. DOI: *https://doi.org/10.5281/zenodo.3785703*

**Koven, CD,** et al. 2020. 'Benchmarking and parameter sensitivity of physiological and vegetation dynamics using the Functionally Assembled Terrestrial Ecosystem Simulator (FATES) at Barro Colorado Island, Panama'. *Biogeosciences*, 17(11): 3017–3044. DOI: *https://doi.org/10.5194/bg-17-3017-2020*

**Lawrence, DM,** et al. 2019. 'The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty'. *Journal of Advances in Modeling Earth Systems*, 11(12): 4245–4287. DOI: *https://doi.org/10.1029/2018MS001583*

**Longo, M,** et al. 2019. 'The biophysics, ecology, and biogeochemistry of functionally diverse, vertically and horizontally heterogeneous ecosystems: the Ecosystem Demography model, version 2.2 – Part 1: Model description'. *Geoscientific Model Development*, 12(10): 4309–4346. DOI: *https://doi.org/10.5194/gmd-12-4309-2019*

**Markstrom, SL,** et al. 2015. *PRMS-IV, the precipitation-runoff modeling system, version* 4, *PRMS-IV, the precipitation-runoff modeling system, version* 4. USGS Numbered Series 6-B7. Reston, VA: U.S. Geological Survey, 169. DOI: *https://doi.org/10.3133/tm6B7*

**McGuire, AD,** et al. 2018. 'Dependence of the evolution of carbon dynamics in the northern permafrost region on the trajectory of climate change'. *Proceedings of the National Academy of Sciences*, 115(15): 3882–3887. DOI: *https://doi.org/10.1073/pnas.1719903115*

**Mekonnen, ZA,** et al. 2019 'Expansion of high-latitude deciduous forests driven by interactions between climate warming and fire'. *Nature Plants*, 5(9): 952–958. DOI: *https://doi.org/10.1038/s41477-019-0495-8*

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

**Phillips, TJ,** et al. 2017. 'Using ARM Observations to Evaluate Climate Model Simulations of Land-Atmosphere Coupling on the U.S. Southern Great Plains'. *Journal of Geophysical Research: Atmospheres*, 122(21): 11,524–11,548. DOI: *https://doi.org/10.1002/2017JD027141*

**Riley, WJ,** et al. 2021. 'Non-growing season plant nutrient uptake controls Arctic tundra vegetation composition under future climate'. 16(7): 074047. DOI: *https://doi.org/10.1088/1748-9326/ac0e63*

**Riley, WJ, Zhu, Q** and **Tang, JY.** 2018. 'Weaker land–climate feedbacks from nutrient uptake during photosynthesis-inactive periods'. *Nature Climate Change*, 8(11): 1002–1006. DOI: *https://doi.org/10.1038/s41558-018-0325-4*

**Sansone, S-A,** et al. 2019. 'FAIRsharing as a community approach to standards, repositories and policies'. *Nature Biotechnology*, 37(4): 358–367. DOI: *https://doi.org/10.1038/s41587-019-0080-8*

**Simmonds, MB, Riley, WJ, Agarwal, DA, Chen, X, Cholia, S, Crystal-Ornelas, R, Coon, ET, Dwivedi, D, Huang, M, Jan, A, Kakalia, Z, Kumar, J, Koven, CD, Li, L, Melara, M, Ricciuto, DM, Walker, AP, Zhi, W, Zhu, Q** and **Varadharajan, C.** 2021. ESS-DIVE guidelines for archiving terrestrial model data. *Environmental Systems Science Data Infrastructure for a Virtual Ecosystem, ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1813868* on 2021-11-16. DOI: *https://doi.org/10.15485/1813868*

**Smith, B, Prentice, IC** and **Sykes, MT.** 2001. 'Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space'. *Global Ecology and Biogeography*, 10(6): 621–637. DOI: *https://doi.org/10.1046/j.1466-822X.2001.t01-1-00256.x*

**Smith, M,** et al. 2013. 'The distributed model intercomparison project – Phase 2: Experiment design and summary results of the western basin experiments'. *Journal of Hydrology*, 507: 300–329. DOI: *https://doi.org/10.1016/j.jhydrol.2013.08.040*

**Sood, A** and **Smakhtin, V.** 2015. 'Global hydrological models: a review'. *Hydrological Sciences Journal*, 60(4), 549–565. DOI: *https://doi.org/10.1080/02626667.2014.950580*

**Stall, S,** et al. 2019. 'Make scientific data FAIR'. *Nature*, 570(7759): 27–29. DOI: *https://doi.org/10.1038/d41586-019-01720-7*

**Steefel, CI** and **Molins, S.** 2009. 'CrunchFlow'. *Software for modeling multicomponent reactive flow and transport. User's manual*. Berkeley: Lawrence Berkeley National Laboratory [Preprint].

**Varadharajan, C,** et al. 2019. 'Launching an Accessible Archive of Environmental Data'. *Eos*. DOI: *https://doi.org/10.1029/2019EO111263*

**Velliquette, T, Welch, J, Crow, M, Devarakonda, R, Heinz, S** and **Crystal-Ornelas, R.** 2021. ESS-DIVE Reporting Format for File-level Metadata. *Environmental Systems Science Data Infrastructure for a Virtual Ecosystem, ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1734840* on 2021-11-16. DOI: *https://doi.org/10.15485/1734840*

**Vittorio, AD** and **Simmonds, M.** 2019. aldivi/caland: CALAND v3.0.0. *Zenodo*. DOI: *https://doi.org/10.5281/zenodo.3256727*

**Walker, AP,** et al. 2014. 'Comprehensive ecosystem model-data synthesis using multiple data sets at two temperate forest free-air $CO_2$ enrichment experiments: Model performance at ambient $CO_2$ concentration'. *Journal of Geophysical Research: Biogeosciences*, 119(5): 937–964. DOI: *https://doi.org/10.1002/2013JG002553*

**Walker, AP,** et al. 2019. 'Decadal biomass increment in early secondary succession woody ecosystems is increased by $CO_2$ enrichment'. *Nature Communications*, 10(1): p. 454. DOI: *https://doi.org/10.1038/s41467-019-08348-1*

**Walker, AP, De Kauwe, MG, Medlyn, B, Zaehle, S, Asao, S, Guenet, B, Harper, A, Hickler, T, Jain, AK, Luo, Y, Lu, X, Luus, K, Shu, S, Wang, Y, Werner, C, Xia, J** and **Norby, RJ.** 2018. FACE-MDS Phase 2: Model Output. Free Air $CO_2$ Enrichment Model Data Synthesis (FACE-MDS). *ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1480327* on 2021-11-15. DOI: *https://doi.org/10.1038/s41467-019-08348-1*

**Walker, AP, Yang, B, Boden, T, De Kauwe, MG, Fenstermaker, LF, Medlyn, B, Megonigal, JP, Oren, R, Pendall, E, Zak, DR, Zaehle, S, Burton, AJ, Drake, BG, Evans, RD, Hungate, B, Johnson, DP, Kim, D, LeCain, D, Lewin, KF, Lu, M, Mueller, KF, Nowak, RS, Riggs, JS, Smith, SD, Tharp, LM, Zelikova, TJ** and **Norby, RJ.** 2018. FACE-MDS Phase 2: Meteorological Data and Protocols. Free Air $CO_2$ Enrichment Model Data Synthesis (FACE-MDS). *ESS-DIVE repository. Dataset*. Accessed via *https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1480328* on 2021-11-16. DOI: *https://doi.org/10.15485/1480328*

**Walker, AP** and **Ye, M,** et al. 2018. 'The multi-assumption architecture and testbed (MAAT v1.0): R code for generating ensembles with dynamic model structure and analysis of epistemic uncertainty from multiple sources'. *Geoscientific Model Development*, 11(8): 3159–3185. DOI: *https://doi.org/10.5194/gmd-11-3159-2018*

**Wilkinson, MD,** et al. 2016. 'The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific Data*, 3: p. 160018. DOI: *https://doi.org/10.1038/sdata.2016.18*

**Woodward, FI** and **Lomas, MR.** 2004. 'Vegetation dynamics – simulating responses to climatic change'. *Biological Reviews*, 79(3): 643–670. DOI: *https://doi.org/10.1017/S1464793103006419*

**Zhi, W,** et al. (2019) 'Distinct Source Water Chemistry Shapes Contrasting Concentration-Discharge Patterns'. *Water Resources Research*, 55(5): 4233–4251. DOI: *https://doi.org/10.1029/2018WR024257*

**Zhu, B,** et al. 2020. 'Effects of Irrigation on Water, Carbon, and Nitrogen Budgets in a Semiarid Watershed in the Pacific Northwest: A Modeling Study'. *Journal of Advances in Modeling Earth Systems*, 12(9): e2019MS001953. DOI: *https://doi.org/10.1029/2019MS001953*

**Zhu, B,** et al. 2021. 'Impact of Vegetation Physiology and Phenology on Watershed Hydrology in a Semiarid Watershed in the Pacific Northwest in a Changing Climate'. *Water Resources Research*, 57(3): e2020WR028394. DOI: *https://doi.org/10.1029/2020WR028394*

**Zhu, Q, Riley, WJ** and **Tang, J.** 2017. 'A new theory of plant–microbe nutrient competition resolves inconsistencies between observations and model predictions'. *Ecological Applications*, 27(3): 875–886. DOI: *https://doi.org/10.1002/eap.1490*

TO CITE THIS ARTICLE:
Simmonds, MB, Riley, WJ, Agarwal, DA, Chen, X, Cholia, S, Crystal-Ornelas, R, Coon, ET, Dwivedi, D, Hendrix, VC, Huang, M, Jan, A, Kakalia, Z, Kumar, J, Koven, CD, Li, L, Melara, M, Ramakrishnan, L, Ricciuto, DM, Walker, AP, Zhi, W, Zhu, Q and Varadharajan, C. 2022. Guidelines for Publicly Archiving Terrestrial Model Data to Enhance Usability, Intercomparison, and Synthesis. *Data Science Journal*, 21: 3, pp. 1–18. DOI: *https://doi.org/10.5334/dsj-2022-003*

Submitted: 22 June 2021
Accepted: 23 November 2021
Published: 07 February 2022



*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.