Publish your terrestrial model data with ESS-DIVE

Charuleka Varadharajan

ESS-DIVE Deputy Lead, Lead: Community Engagement

March 2022 Webinar

Objectives

- What are model data?
- ESS-DIVE's community-oriented process for model-data archiving
- New guidelines for storing model data with ESS-DIVE
- Other ESS-DIVE capabilities for hosting model data

- **Feedback on additional needs** for publishing model data on ESS-DIVE.

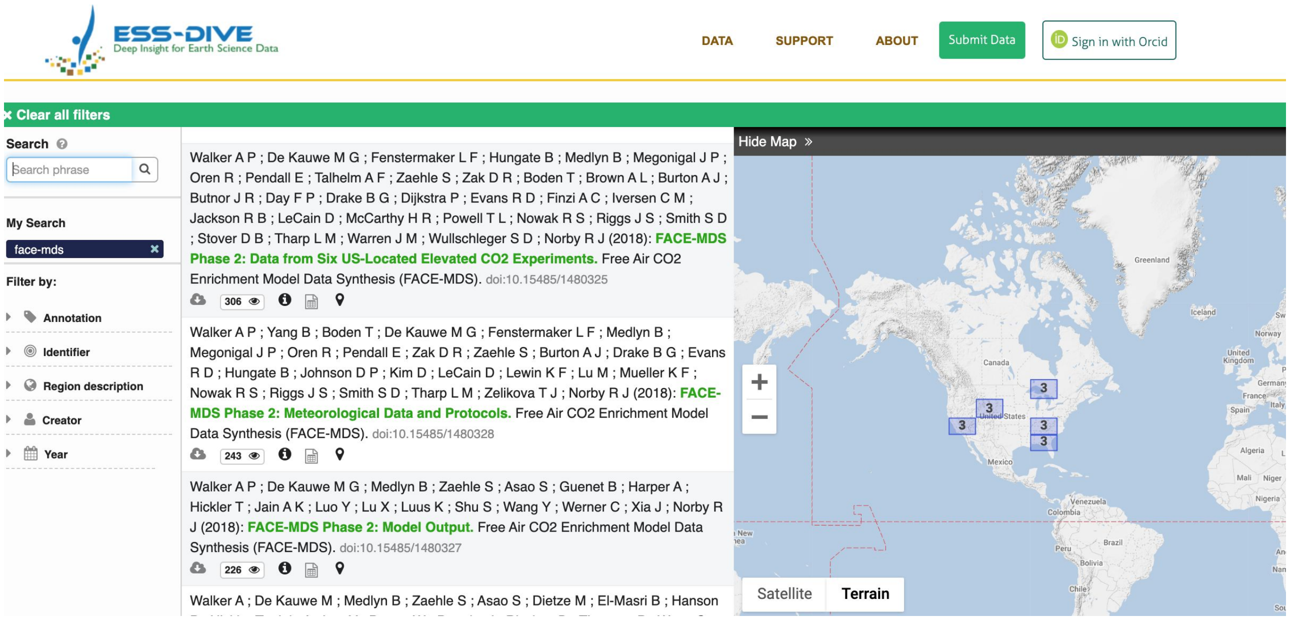
Learn about ESS-DIVE's new guidelines for archiving model data and apply it to your datasets!

What does the "model data" mean to you?

What are “model data”?

- Model outputs: Various dimensions and resolutions of final raw output files, spin-up output files, restart files, test data files, higher level output files corresponding to figures or tables
- Model input files: e.g. forcings, parameters
- Metadata files
- Model code
- Scripts for model set-up and initialization; parameterization; post-processing; and visualizations.
- Visualization files

Model data on ESS-DIVE



Walker A P ; De Kauwe M G ; Fenstermaker L F ; Hungate B ; Medlyn B ; Megonigal J P ; Oren R ; Pendall E ; Talhelm A F ; Zaehle S ; Zak D R ; Boden T ; Brown A L ; Burton A J ; Butnor J R ; Day F P ; Drake B G ; Dijkstra P ; Evans R D ; Finzi A C ; Iversen C M ; Jackson R B ; LeCain D ; McCarthy H R ; Powell T L ; Nowak R S ; Riggs J S ; Smith S D ; Stover D B ; Tharp L M ; Warren J M ; Wullschleger S D ; Norby R J (2018): **FACE-MDS Phase 2: Data from Six US-Located Elevated CO2 Experiments.** Free Air CO2 Enrichment Model Data Synthesis (FACE-MDS). doi:10.15485/1480325

306

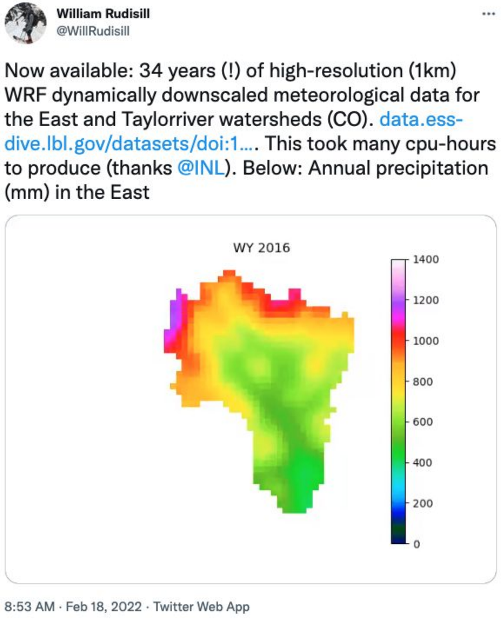
Walker A P ; Yang B ; Boden T ; De Kauwe M G ; Fenstermaker L F ; Medlyn B ; Megonigal J P ; Oren R ; Pendall E ; Zak D R ; Zaehle S ; Burton A J ; Drake B G ; Evans R D ; Hungate B ; Johnson D P ; Kim D ; LeCain D ; Lewin K F ; Lu M ; Mueller K F ; Nowak R S ; Riggs J S ; Smith S D ; Tharp L M ; Zelikova T J ; Norby R J (2018): **FACE-MDS Phase 2: Meteorological Data and Protocols.** Free Air CO2 Enrichment Model Data Synthesis (FACE-MDS). doi:10.15485/1480328

243

Walker A P ; De Kauwe M G ; Medlyn B ; Zaehle S ; Asao S ; Guenet B ; Harper A ; Hickler T ; Jain A K ; Luo Y ; Lu X ; Luus K ; Shu S ; Wang Y ; Werner C ; Xia J ; Norby R J (2018): **FACE-MDS Phase 2: Model Output.** Free Air CO2 Enrichment Model Data Synthesis (FACE-MDS). doi:10.15485/1480327

226

Walker A ; De Kauwe M ; Medlyn B ; Zaehle S ; Asao S ; Dietze M ; El-Masri B ; Hanson



William Rudisill
@WillRudisill

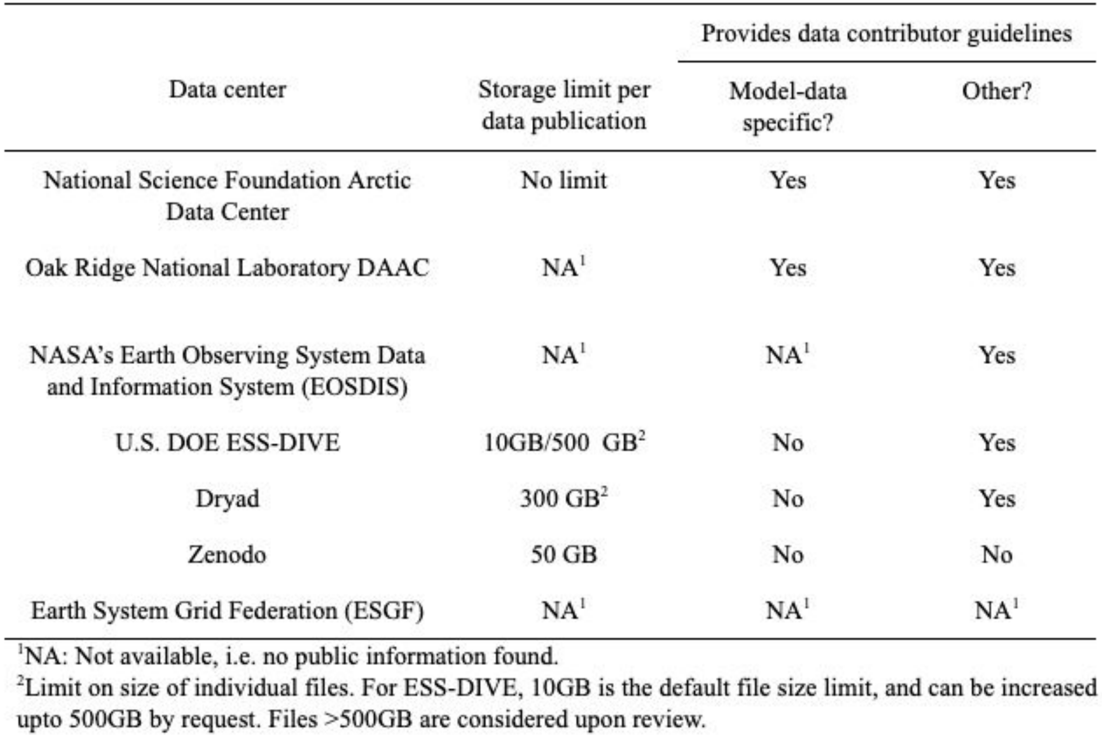
Now available: 34 years (!) of high-resolution (1km) WRF dynamically downscaled meteorological data for the East and Taylorriver watersheds (CO). data.ess-dive.lbl.gov/datasets/doi:1…. This took many cpu-hours to produce (thanks @INL). Below: Annual precipitation (mm) in the East

8:53 AM · Feb 18, 2022 · Twitter Web App

Rudisill et al:
https://data.ess-dive.lbl.gov/view/doi:10.15485/1845448

Process to develop guidelines for terrestrial model data archiving

Review of data systems that store model data



		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]

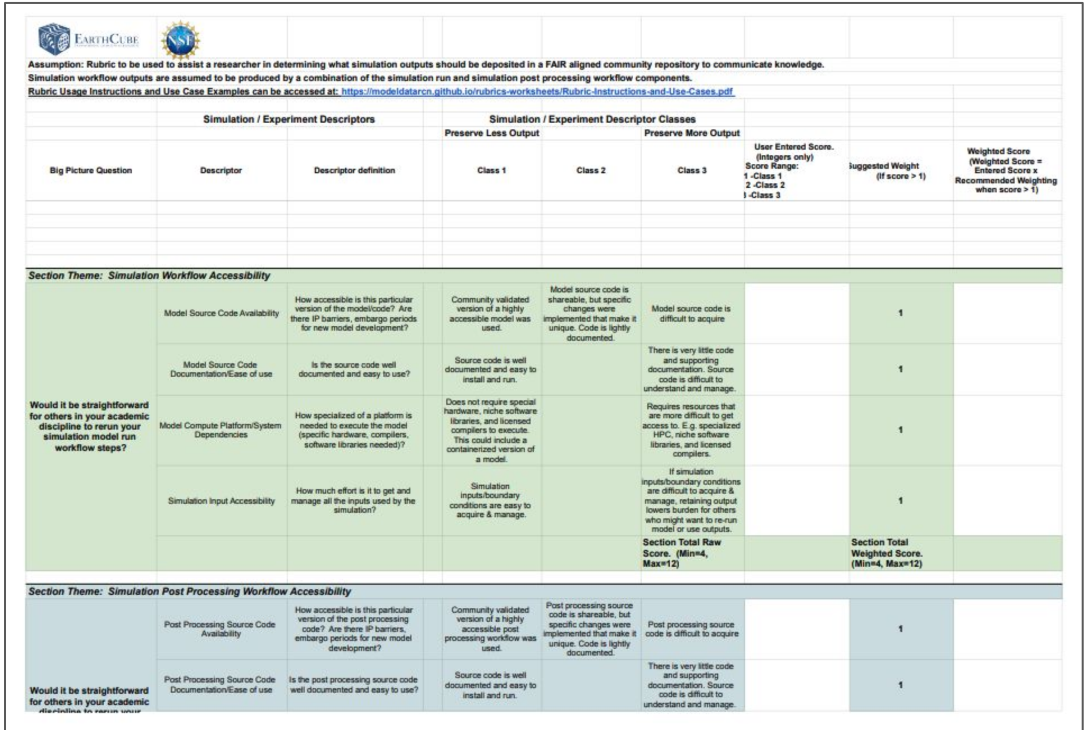
[1] NA: Not available, i.e. no public information found.

[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.

Simmonds et al., 2022, Data Sci. Journ.

Other community guidelines

NSF Earthcube Model Data Rubric/ AGU guidelines



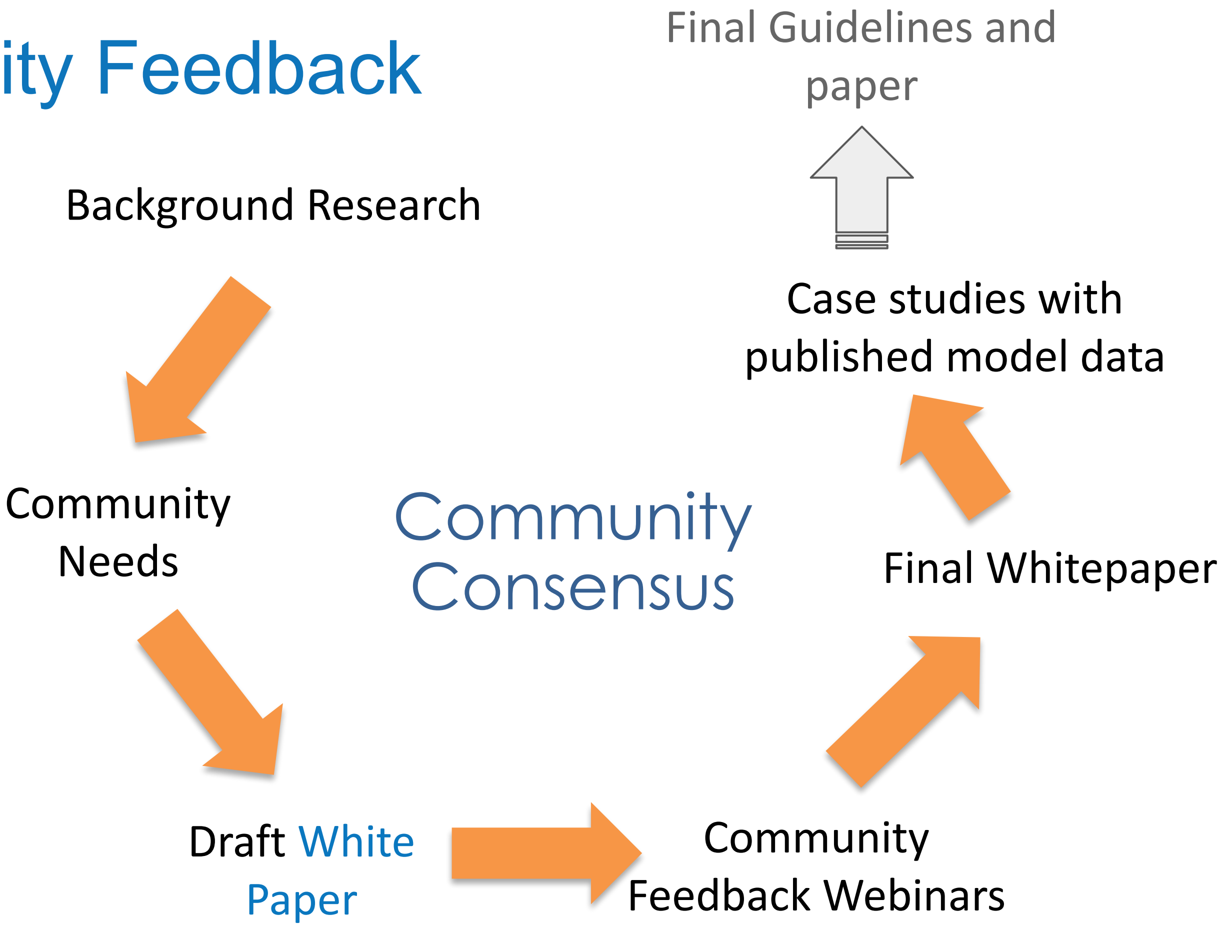
Assumption: Rubric to be used to assist a researcher in determining what simulation outputs should be deposited in a FAIR aligned community repository to communicate knowledge.

Simulation workflow outputs are assumed to be produced by a combination of the simulation run and simulation post processing workflow components.

Rubric Usage Instructions and Use Case Examples can be accessed at: https://modeldatarcn.github.io/rubrics-worksheets/Rubric-Instructions-and-Use-Cases.pdf

	Simulation / Experiment Descriptors		Simulation / Experiment Descriptor Classes					
			Preserve Less Output		Preserve More Output			
Big Picture Question	Descriptor	Descriptor definition	Class 1	Class 2	Class 3	User Entered Score. (Integers only) Score Range: 1 -Class 1 2 -Class 2 3 -Class 3	Suggested Weight (If score > 1)	Weighted Score (Weighted Score = Entered Score x Recommended Weighting when score > 1)
Section Theme: Simulation Workflow Accessibility								
Would it be straightforward for others in your academic discipline to rerun your simulation model run workflow steps?	Model Source Code Availability	How accessible is this particular version of the model/code? Are there IP barriers, embargo periods for new model development?	Community validated version of a highly accessible model was used.	Model source code is shareable, but specific changes were implemented that make it unique. Code is lightly documented.	Model source code is difficult to acquire		1	
	Model Source Code Documentation/Ease of use	Is the source code well documented and easy to use?	Source code is well documented and easy to install and run.		There is very little code and supporting documentation. Source code is difficult to understand and manage.		1	
	Model Compute Platform/System Dependencies	How specialized of a platform is needed to execute the model (specific hardware, compilers, software libraries needed)?	Does not require special hardware, niche software libraries, and licensed compilers to execute. This could include a containerized version of a model.		Requires resources that are more difficult to get access to. E.g. specialized HPC, niche software libraries, and licensed compilers.		1	
	Simulation Input Accessibility	How much effort is it to get and manage all the inputs used by the simulation?	Simulation inputs/boundary conditions are easy to acquire & manage.		If simulation inputs/boundary conditions are difficult to acquire & manage, retaining output lowers burden for others who might want to re-run model or use outputs.		1	
					Section Total Raw Score. (Min=4, Max=12)		**Section Total Weighted Score. (Min=4, Max=12)**	
Section Theme: Simulation Post Processing Workflow Accessibility								
Would it be straightforward for others in your academic discipline to rerun your	Post Processing Source Code Availability	How accessible is this particular version of the post processing code? Are there IP barriers, embargo periods for new model development?	Community validated version of a highly accessible post processing workflow was used.	Post processing source code is shareable, but specific changes were implemented that make it unique. Code is lightly documented.	Post processing source code is difficult to acquire		1	
	Post Processing Source Code Documentation/Ease of use	Is the post processing source code well documented and easy to use?	Source code is well documented and easy to install and run.		There is very little code and supporting documentation. Source code is difficult to understand and manage.		1	

Community Feedback



Background Research

Community Needs

Draft White Paper

Community Feedback Webinars

Final Whitepaper

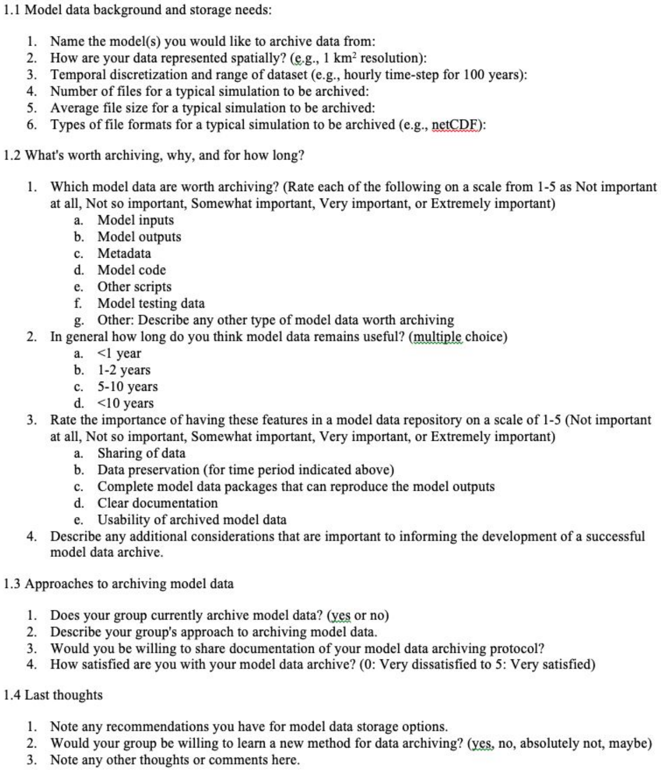
Case studies with published model data

Final Guidelines and paper

Community Consensus

Questions to Modelers

- Gathered information on
  - Models used
  - File storage and specifications
  - What is worth archiving and how long is the data useful?
  - Archiving protocols
  - Features needed from ESS-DIVE
- Circulated widely to ESS community and conducted 2 webinars on this feedback



1.1 Model data background and storage needs:

1. Name the model(s) you would like to archive data from:
2. How are your data represented spatially? (e.g., 1 km^2 resolution):
3. Temporal discretization and range of dataset (e.g., hourly time-step for 100 years):
4. Number of files for a typical simulation to be archived:
5. Average file size for a typical simulation to be archived:
6. Types of file formats for a typical simulation to be archived (e.g., netCDF):

1.2 What's worth archiving, why, and for how long?

1. Which model data are worth archiving? (Rate each of the following on a scale from 1-5 as Not important at all, Not so important, Somewhat important, Very important, or Extremely important)
   a. Model inputs
   b. Model outputs
   c. Metadata
   d. Model code
   e. Other scripts
   f. Model testing data
   g. Other: Describe any other type of model data worth archiving
2. In general how long do you think model data remains useful? (multiple choice)
   a. <1 year
   b. 1-2 years
   c. 5-10 years
   d. <10 years
3. Rate the importance of having these features in a model data repository on a scale of 1-5 (Not important at all, Not so important, Somewhat important, Very important, or Extremely important)
   a. Sharing of data
   b. Data preservation (for time period indicated above)
   c. Complete model data packages that can reproduce the model outputs
   d. Clear documentation
   e. Usability of archived model data
4. Describe any additional considerations that are important to informing the development of a successful model data archive.

1.3 Approaches to archiving model data

1. Does your group currently archive model data? (yes or no)
2. Describe your group's approach to archiving model data.
3. Would you be willing to share documentation of your model data archiving protocol?
4. How satisfied are you with your model data archive? (0: Very dissatisfied to 5: Very satisfied)

1.4 Last thoughts

1. Note any recommendations you have for model data storage options.
2. Would your group be willing to learn a new method for data archiving? (yes, no, absolutely not, maybe)
3. Note any other thoughts or comments here.

Participants and projects involved

Researcher	DOE Project(s) Represented
Charlie Koven	NGEE-Tropics, NGEE-Arctic
Jitu Kumar	NGEE-Tropics, NGEE-Arctic
Dipankar Dwivedi	Watershed Function SFA
Anthony Walker	NGEE-Tropics, FACE-MDS, Oakridge TES SFA, RUBISCO, NGEE-Arctic
Xingyuan Chen	PNNL SBR SFA, IDEAS, EXOSHEDS
Scott Painter	NGEE-Arctic, IDEAS, Oakridge Mercury SFA, Exosheds
Dan Ricciuto	Oakridge TES SFA
Qing Zhu	E3SM, RUBISCO
Ethan Coon	NGEE-Arctic, Exasheds
Maoyi Hung	NGEE Tropics, SBC SFA
Kate Maher	SLAC SFA
Ahmad Jan	NGEE-Arctic

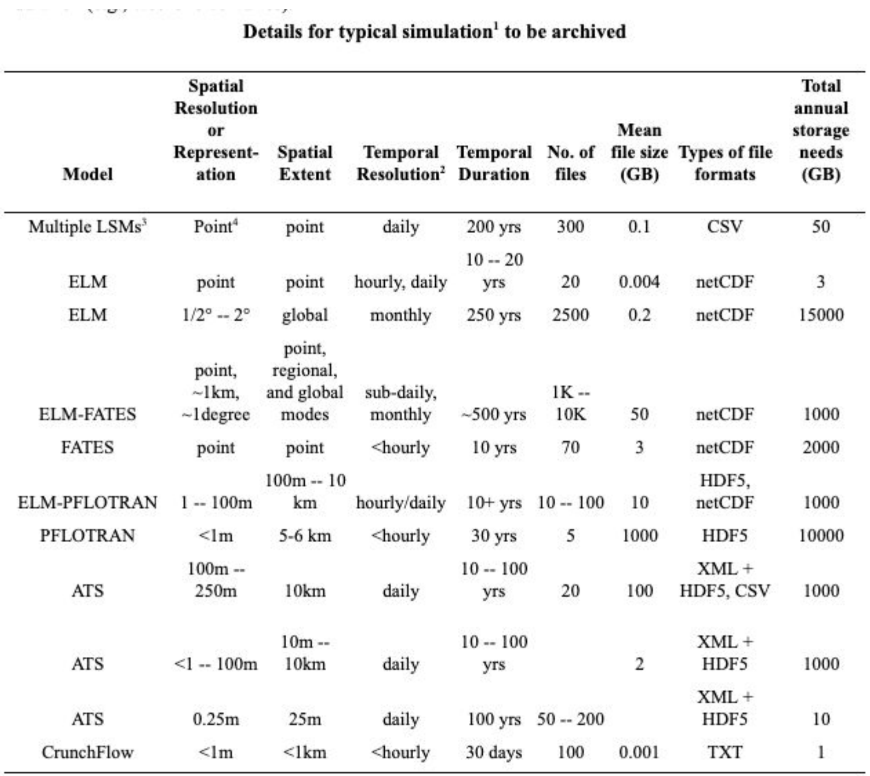
Models used across DOE projects

- ELM
- ELM-FATES
- ELM-BeTR
- CLM
- PFLOTRAN
- CABLE
- SDVGM
- GDAY
- ED2
- LPJ-GUESS

- MAAT
- ATF
- ATS
- OpenFOAM (CFD)
- Crunch
- Crunch-Flow
- SWAT
- Machine Learning

File storage and specifications

- Total files per simulation: 5 to a few million files
- Average file size: 100 MB - 2 TB
- Total current storage: 100’s of MB to 100’s of TB (mean = 28 TB/modeler, median = 650 GB/modeler).



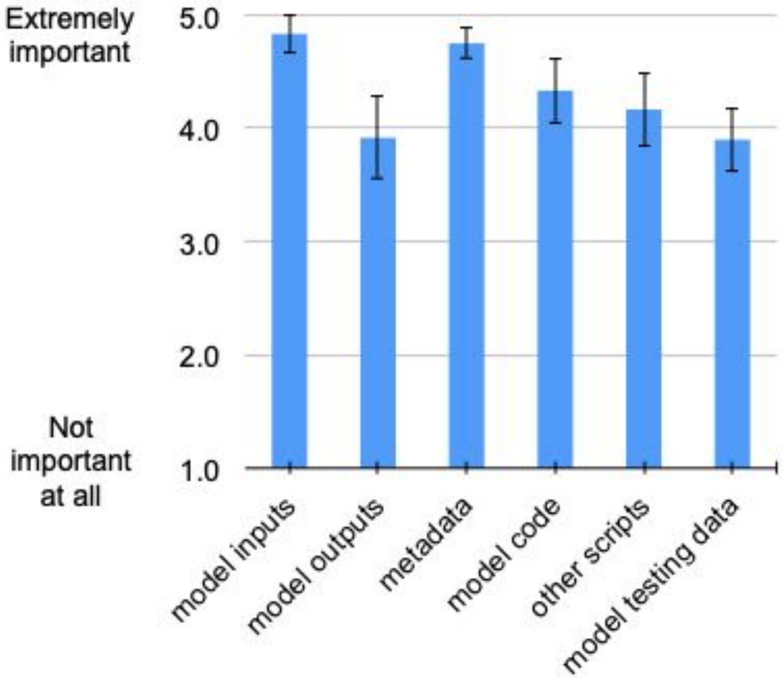
Details for typical simulation[1] to be archived

Model	Spatial Resolution or Represent-ation	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, ~1km, ~1degree	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 -- 100m	100m -- 10 km	hourly/daily	10+ yrs	10 -- 100	10	HDF5, netCDF	1000
PFLOTRAN	<1m	5-6 km	<hourly	30 yrs	5	1000	HDF5	10000
ATS	100m -- 250m	10km	daily	10 -- 100 yrs	20	100	XML + HDF5, CSV	1000
ATS	<1 -- 100m	10m -- 10km	daily	10 -- 100 yrs		2	XML + HDF5	1000
ATS	0.25m	25m	daily	100 yrs	50 -- 200		XML + HDF5	10
CrunchFlow	<1m	<1km	<hourly	30 days	100	0.001	TXT	1

Simmonds et al., 2022, Data Sci. Journ.

What model data components are worth archiving?

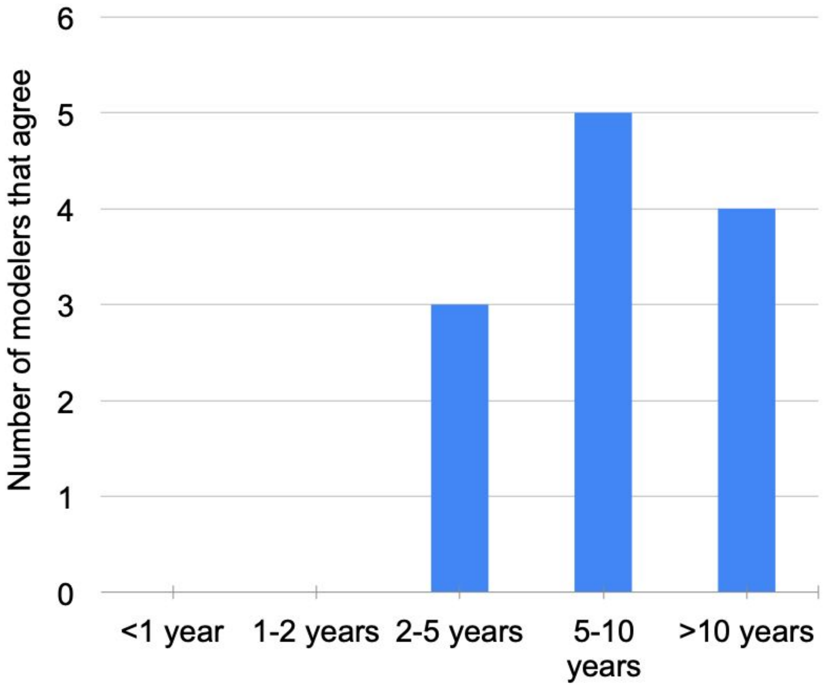
- Journal and funding requirements for data archiving
- Archive almost everything!
- Exception outputs due to size: archive only high-level outputs corresponding to key figures/findings
- Ambiguity around model code



+ restart files

Simmonds et al., 2022, Data Sci. Journ.

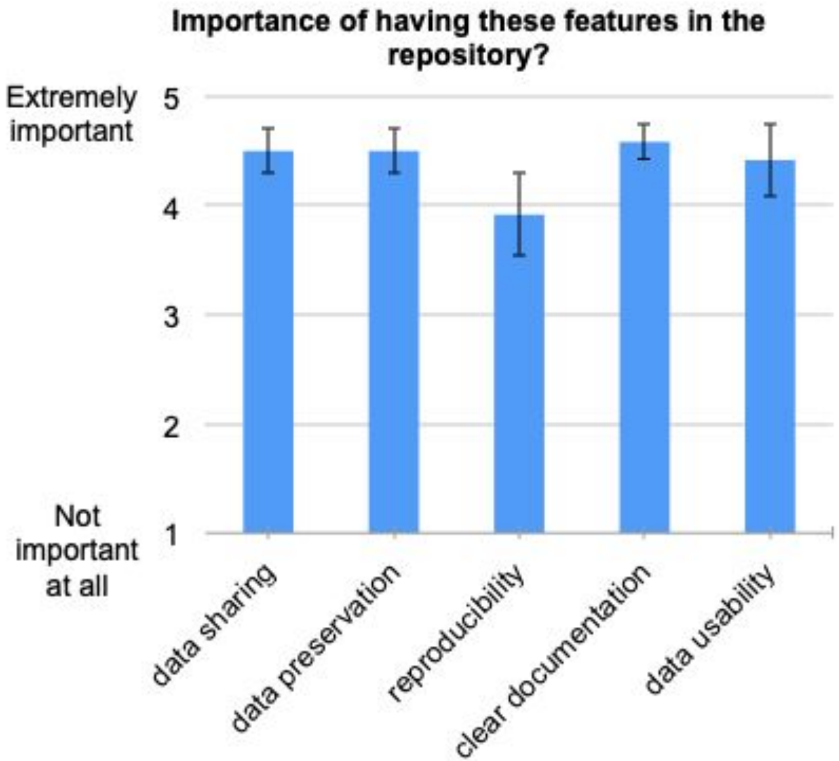
How long does model data remain useful?



Simmonds et al., 2022, Data Sci. Journ.

Purposes for storing model data publicly

- Transparency and reproducibility
- Model intercomparisons and synthesis
- Comparison with observations
- Standards for ensuring data reusability
  - Machine-readable
  - Critical metadata
  - Workflow documents
  - 90% willing to learn new guidelines or standardized reporting format for model data

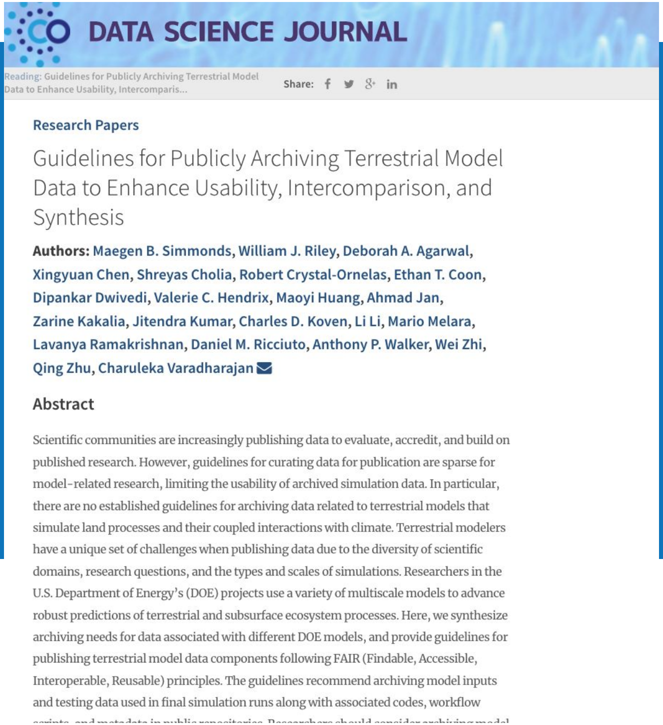


Simmonds et al., 2022, Data Sci. Journ.

Any Questions?

Photo by Towfiqu barbhuiya on Unsplash

Guidelines:
How to archive model data


DATA SCIENCE JOURNAL
Reading: Guidelines for Publicly Archiving Terrestrial Model Data to Enhance Usability, Intercomparis...
Share:

Research Papers

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

Authors: 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 ✉

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

Components of model data

- Metadata
- Required Data Files
  - Model Inputs
  - Model outputs
  - Model code
  - Scripts
- Optional Files
  - File-level metadata (FLMD)
  - Model Testing/Validation Data
  - Documentation or user guide

Model Input Files

- Input files required unless publicly available elsewhere
  - Examples: climate forcings, meshes, soil parameterizations
  - Use open-sourced formats such as comma separated value (.csv) or NetCDF (.nc) formats where possible
  - File names should be unique, should only contain letters, numbers, hyphens, underscores, should not contain spaces, and should not rely on case-sensitive file systems
  - Hyperlink to specific input files in metadata and user guide

Use external linking feature if input files are publicly available on another established repository

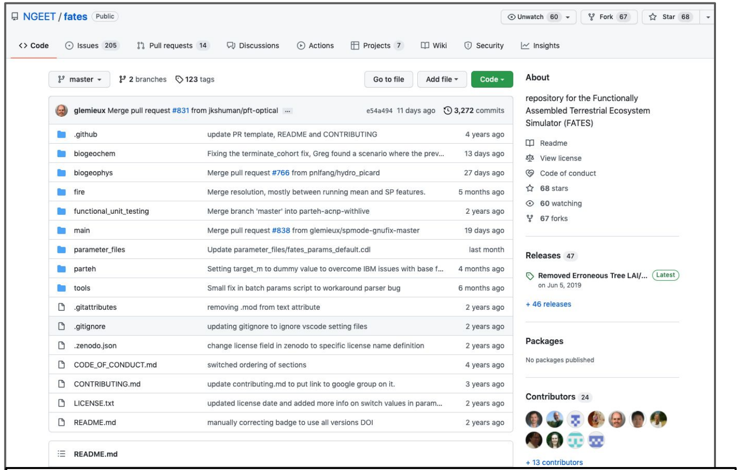
Model output files

- Includes raw and post-processed data, data supporting findings, tables, figures in a paper.
- Archive all model outputs if the size of the data files are within the repository **storage limitations (500GB/file on ESS-DIVE)**
- Use decision tree if size of model output files exceed repository storage limits
- Open-sourced formats such as comma separated value (.csv) or NetCDF (.nc) formats where possible.

Contact ess-dive-support@lbl.gov if you have datasets > 0.5 TB

Software

- Model Code
  - Include source code(s) unless publicly available
  - Include specific version, commit hash, or citation allowing exact source code to be recovered
  - If available, include external link to tagged release in an established data repository.
  - Include links to model codes in metadata/user guide
- Scripts
  - Include **run scripts** if necessary for running model to generate results
  - Optional: Scripts for reproducing parameters and model configuration for simulations and input files, post-processing model outputs, for executing entire workflow.



glemieux Merge pull request #831 from jkshuman/pft-optical … e54a494 11 days ago 3,272 commits

.github	update PR template, README and CONTRIBUTING	4 years ago
biogeochem	Fixing the terminate_cohort fix, Greg found a scenario where the prev…	13 days ago
biogeophys	Merge pull request #766 from pnlfang/hydro_picard	27 days ago
fire	Merge resolution, mostly between running mean and SP features.	5 months ago
functional_unit_testing	Merge branch 'master' into parteh-acnp-withlive	2 years ago
main	Merge pull request #838 from glemieux/spmode-gnufix-master	19 days ago
parameter_files	Update parameter_files/fates_params_default.cdl	last month
parteh	Setting target_m to dummy value to overcome IBM issues with base f…	4 months ago
tools	Small fix in batch params script to workaround parser bug	6 months ago
.gitattributes	removing .mod from text attribute	2 years ago
.gitignore	updating gitignore to ignore vscode setting files	2 years ago
.zenodo.json	change license field in zenodo to specific license name definition	2 years ago
CODE_OF_CONDUCT.md	switched ordering of sections	4 years ago
CONTRIBUTING.md	update contributing.md to put link to google group on it.	3 years ago
LICENSE.txt	updated license date and added more info on switch values in param…	2 years ago
README.md	manually correcting badge to use all versions DOI	2 years ago

README.md

About

repository for the Functionally Assembled Terrestrial Ecosystem Simulator (FATES)

Readme
View license
Code of conduct
68 stars
60 watching
67 forks

Releases 47

Removed Erroneous Tree LAI/… Latest
on Jun 5, 2019

+ 46 releases

Packages

No packages published

Contributors 24

+ 13 contributors

Codes published within an ESS-DIVE dataset are CC0 or CCBy4 license. Use external linking feature with software on DOE CODE

Optional Elements

- File-Level Metadata
- Model Testing/Validation Data
  - Data files of observations from each location in an open source format (e.g., CSV).
  - If data are publicly available in another repository, use external linking and include a reference (with DOI) in metadata and user guide.
- Documentation or user guide
  - Readme file (e.g., pdf) for each site-specific or large-scale simulation
  - Provide details on model name, version number, and required data or code dependencies.
  - Include a citation for the model code and licensing information if applicable.

File-level metadata

What are file-level metadata?

- Granular information at the data file level (e.g., file name & description, start and end dates)

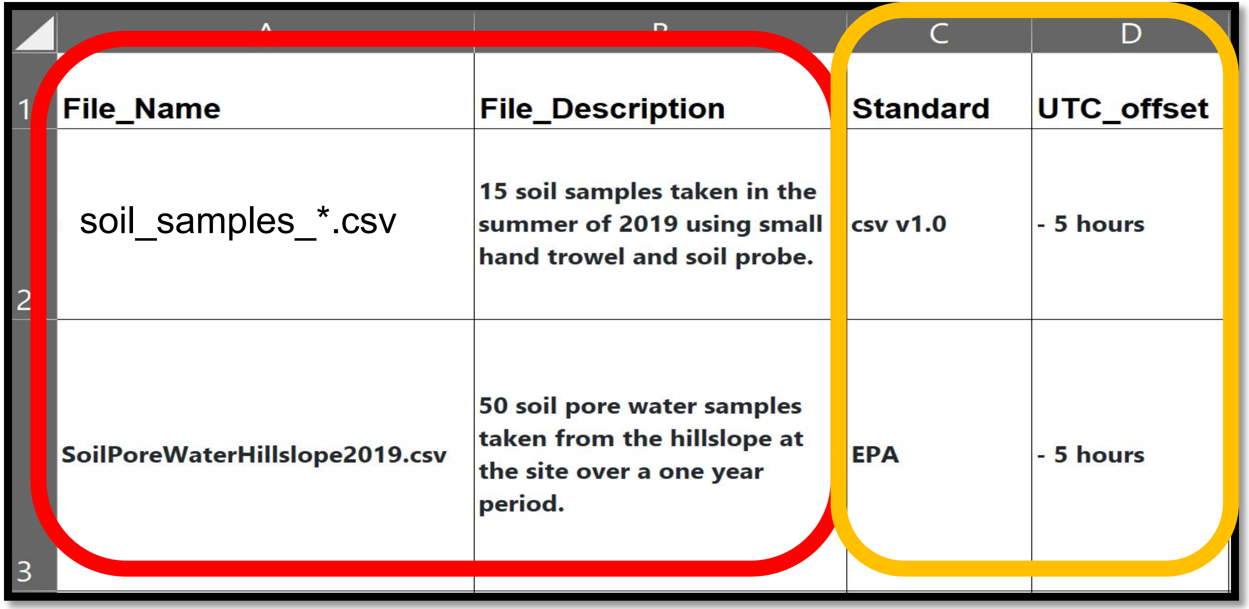
Why provide file-level metadata?

- Data users will have general understanding of info contained within a file
- FLMD can enable automatic parsing of data files so that users can eventually search & locate files across data collections

Terri Velliquette

OAK RIDGE National Laboratory

File-level metadata example



File_Name	File_Description	Standard	UTC_offset
soil_samples_*.csv	15 soil samples taken in the summer of 2019 using small hand trowel and soil probe.	csv v1.0	- 5 hours
SoilPoreWaterHillslope2019.csv	50 soil pore water samples taken from the hillslope at the site over a one year period.	EPA	- 5 hours

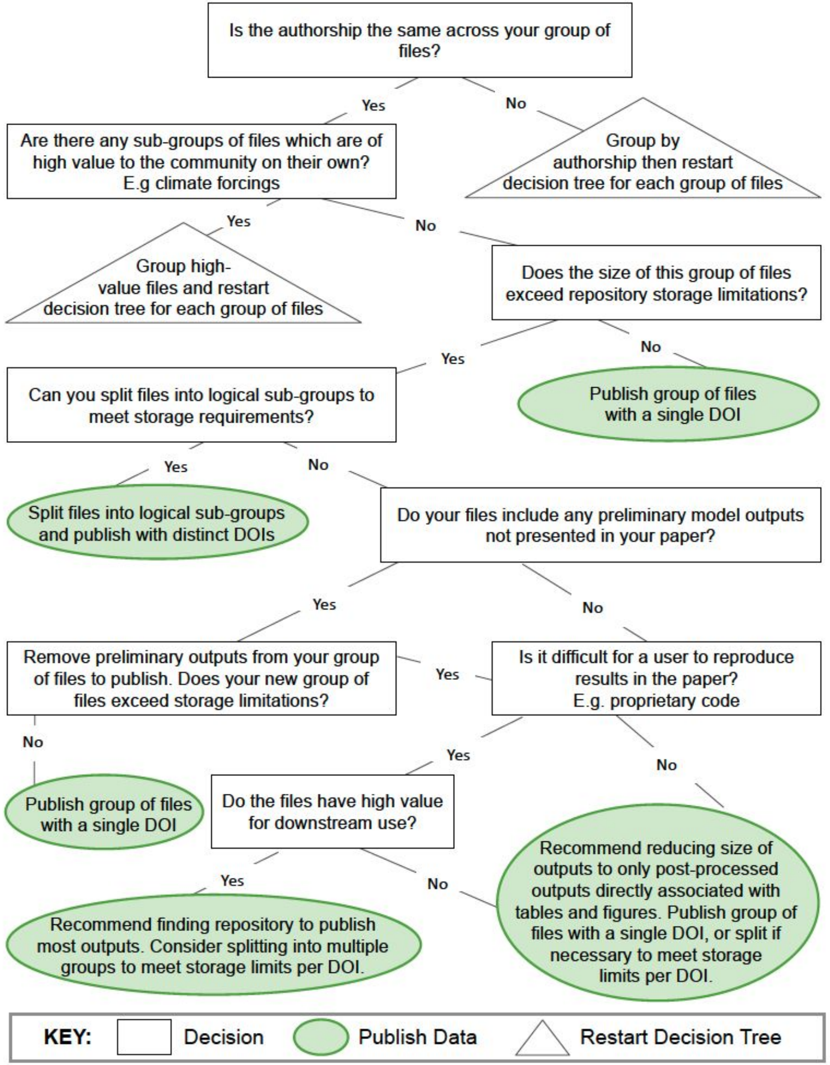
- FLMD **template**: https://ess-dive.gitbook.io/file-level-metadata-reporting-format/
- Can use wildcard * to indicate when FLMD applies to multiple files

Data package guidelines

Key Criteria

- Authorship
- File size/Repository storage capacity
- Decide what model outputs are worth archiving

Simmonds et al., 2022, Data Sci. Journ.



Is the authorship the same across your group of files?

Yes → Are there any sub-groups of files which are of high value to the community on their own? E.g climate forcings

No → Group by authorship then restart decision tree for each group of files

Are there any sub-groups… Yes → Group high-value files and restart decision tree for each group of files

Are there any sub-groups… No → Does the size of this group of files exceed repository storage limitations?

Does the size… No → Publish group of files with a single DOI

Does the size… Yes → Can you split files into logical sub-groups to meet storage requirements?

Can you split… Yes → Split files into logical sub-groups and publish with distinct DOIs

Can you split… No → Do your files include any preliminary model outputs not presented in your paper?

Do your files include… Yes → Remove preliminary outputs from your group of files to publish. Does your new group of files exceed storage limitations?

Remove preliminary outputs… No → Publish group of files with a single DOI

Remove preliminary outputs… Yes → Is it difficult for a user to reproduce results in the paper? E.g. proprietary code

Do your files include… No → Is it difficult for a user to reproduce results in the paper? E.g. proprietary code

Is it difficult… Yes → Do the files have high value for downstream use?

Is it difficult… No → Recommend reducing size of outputs to only post-processed outputs directly associated with tables and figures. Publish group of files with a single DOI, or split if necessary to meet storage limits per DOI.

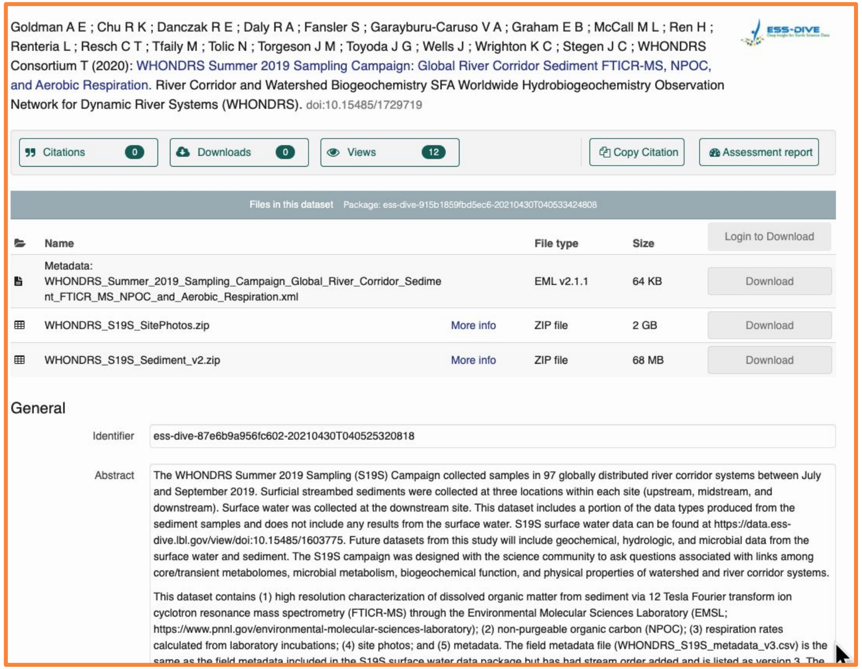
Do the files have high value… Yes → Recommend finding repository to publish most outputs. Consider splitting into multiple groups to meet storage limits per DOI.

Do the files have high value… No → Recommend reducing size of outputs to only post-processed outputs directly associated with tables and figures. Publish group of files with a single DOI, or split if necessary to meet storage limits per DOI.

KEY: □ Decision ● Publish Data △ Restart Decision Tree

Model data in publications

- Cite dataset and include links to the data and code publication(s) in the Data or Code Availability section and references.



Goldman A E ; Chu R K ; Danczak R E ; Daly R A ; Fansler S ; Garayburu-Caruso V A ; Graham E B ; McCall M L ; Ren H ; Renteria L ; Resch C T ; Tfaily M ; Tolic N ; Torgeson J M ; Toyoda J G ; Wells J ; Wrighton K C ; Stegen J C ; WHONDRS Consortium T (2020): WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Sediment FTICR-MS, NPOC, and Aerobic Respiration. River Corridor and Watershed Biogeochemistry SFA Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS). doi:10.15485/1729719

Files in this dataset Package: ess-dive-915b1859fbd5ec6-20210430T040533424808

Name		File type	Size	Login to Download
Metadata: WHONDRS_Summer_2019_Sampling_Campaign_Global_River_Corridor_Sediment_FTICR_MS_NPOC_and_Aerobic_Respiration.xml		EML v2.1.1	64 KB	Download
WHONDRS_S19S_SitePhotos.zip	More info	ZIP file	2 GB	Download
WHONDRS_S19S_Sediment_v2.zip	More info	ZIP file	68 MB	Download

General

Identifier: ess-dive-87e6b9a956fc602-20210430T040525320818

Abstract: The WHONDRS Summer 2019 Sampling (S19S) Campaign collected samples in 97 globally distributed river corridor systems between July and September 2019. Surficial streambed sediments were collected at three locations within each site (upstream, midstream, and downstream). Surface water was collected at the downstream site. This dataset includes a portion of the data types produced from the sediment samples and does not include any results from the surface water. S19S surface water data can be found at https://data.ess-dive.lbl.gov/view/doi:10.15485/1603775. Future datasets from this study will include geochemical, hydrologic, and microbial data from the surface water and sediment. The S19S campaign was designed with the science community to ask questions associated with links among core/transient metabolomes, microbial metabolism, biogeochemical function, and physical properties of watershed and river corridor systems.

This dataset contains (1) high resolution characterization of dissolved organic matter from sediment via 12 Tesla Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) through the Environmental Molecular Sciences Laboratory (EMSL; https://www.pnnl.gov/environmental-molecular-sciences-laboratory); (2) non-purgeable organic carbon (NPOC); (3) respiration rates calculated from laboratory incubations; (4) site photos; and (5) metadata. The field metadata file (WHONDRS_S19S_metadata_v3.csv) is the same as the field metadata included in the S19S surface water data package but has had stream order added and is listed as version 3. The

Example 1: Citing multiple datasets

Walker, AP, et al. 2019. 'Decadal biomass increment in early secondary succession woody ecosystems is increased by CO2 enrichment'. Nature Communications, 10(1): p. 454. DOI: https://doi.org/10.1038/s41467-019-08348-1

The site-based **meteorological dataset** (https://data.ess-dive.lbl.gov/view/ess-dive-7807cf86f1dd42a-20181127T173047368940), the **model output dataset** (https://data.ess-dive.lbl.gov/view/ess-dive-8260043c35fc925-20181130T171955541030) and the **experiment dataset** (https://data.ess-dive.lbl.gov/view/ess-dive-f525c71da7d2681-20181128T160851574946) generated and analyzed during the current study are available at the US Department of Energy's (DOE) ESS-DIVE repository.

Example 2: Citing model code

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

The **FATES model** is available at https://github.com/NGEET/fates (last access: 15 May 2020; https://doi.org/10.5281/zenodo.3825474, FATES Development Team, 2020). Experiments here are based on **git commit 0bc7a5d on the fork: https://github.com/ckoven/fates (last access: 4 June 2020; https://doi.org/10.5281/zenodo.3875687**, FATES Development Team, 2019). FATES is run here within two host land surface models, CLM5 and ELMv1, available at https://github.com/ESCOMP/ctsm (git commit b9c92b7, last access: 15 May 2020; https://doi.org/10.5281/zenodo.3739617, CTSM Development Team, 2020) and https://github.com/E3SM-Project/E3SM (git commit 544db3b, last access: 15 May 2020; https://doi.org/10.11578/E3SM/dc.20180418.36, E3SM Project, 2018), respectively. Scripts to initialize parameter files and analyze model output shown here are available at https://github.com/NGEET/testbeds (last access: 15 May 2020; https://doi.org/10.5281/zenodo.3785705, Koven, 2020a), and scripts to run the all model experiments here are available at https://github.com/ckoven/runscripts (last access: 15 May 2020; https://doi.org/10.5281/zenodo.3785703, Koven, 2020b).

Example 3: Citing data, code, scripts

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

The Advanced Terrestrial Simulator (ATS) (Coon et al., 2019) is open source under the BSD 3-clause license and is publicly available at https://github.com/amanzi/ats (last access: October 2019; Coon, 2016). Simulations were conducted using version 0.88. **The ATS version 0.88 is permanently stored at https://doi.org/10.5281/zenodo.3727209 (Coon et al., 2020)**. **Forcing data, model input files, Jupyter notebooks** used to generate figures, and meshes along with Jupyter notebooks used to generate the meshes are publicly available at https://doi.org/10.5440/1545603 (Jan et al., 2019). **Data products used in the model comparisons are publicly available through the NGEE Arctic long-term data archive** https://doi.org/10.5440/1416559. The observed water level can be accessed at https://doi.org/10.5440/1183767 (Liljedahl and Wilson., 2016), the soil temperature data at https://doi.org/10.5440/1126515 (Romanovsky et al., 2017), and the evapotranspiration data at https://doi.org/10.5440/1362279 (Dengel et al., 2019, respectively.

Any Questions?

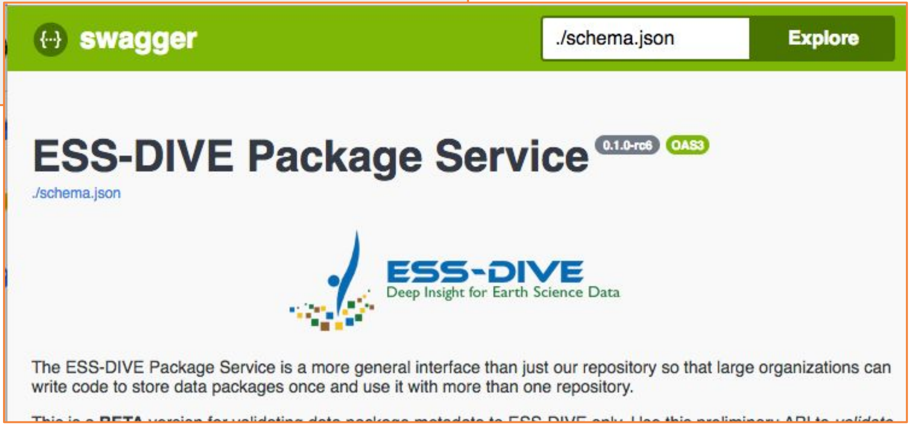
Photo by Towfiqu barbhuiya on Unsplash

ESS-DIVE features and plans to support model data

Application Programming Interface (API) Upload

```
provider_spruce = {
    "name": "SPRUCE",
    "member": {
        "@id": "http://orcid.org/0000-0001-7293-3561 ",
        "givenName": "Paul J",
        "familyName": "Hanson",
        "email": "hansonpj@ornl.gov",
        "jobTitle": "Principal Investigator"
        }
    }
```

Web UI limited to 10GB/file, API allows 500GB/file



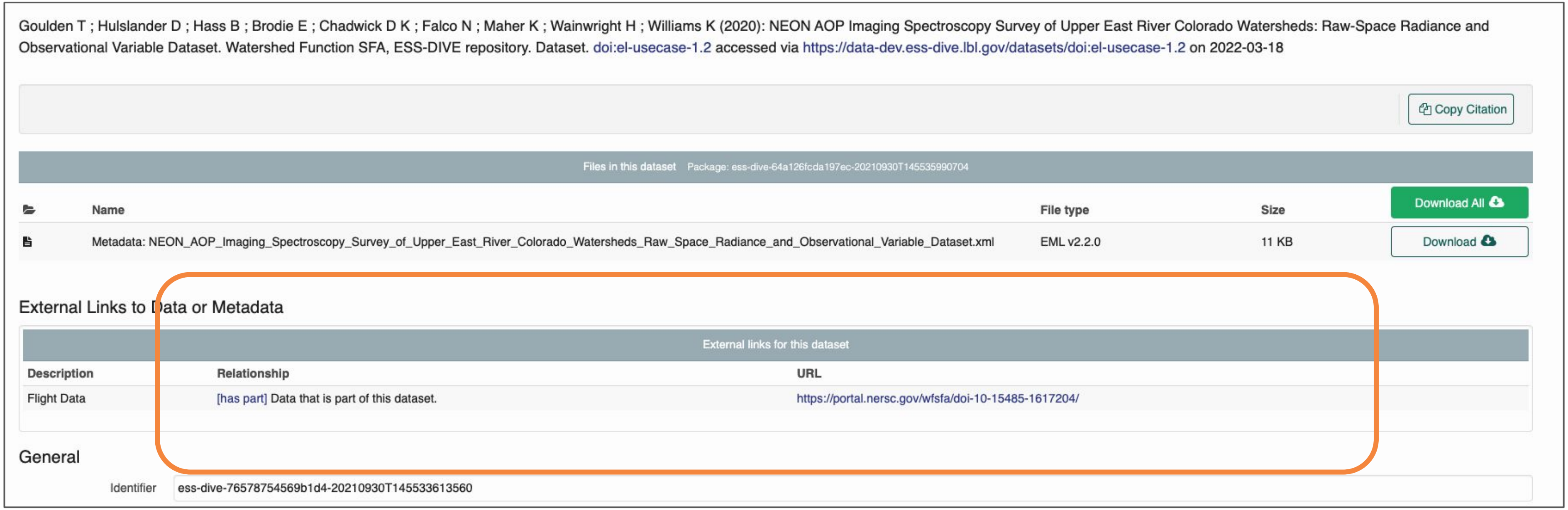
ESS-DIVE Package Service

The ESS-DIVE Package Service is a more general interface than just our repository so that large organizations can write code to store data packages once and use it with more than one repository.

This is a **BETA** version for validating data package metadata to ESS-DIVE only. Use this preliminary API to *validate*

“External Linking” Feature

Data stored on other repositories are directly linked to dataset metadata on ESS-DIVE and **some/all data do not have to be on ESS-DIVE**



Goulden T ; Hulslander D ; Hass B ; Brodie E ; Chadwick D K ; Falco N ; Maher K ; Wainwright H ; Williams K (2020): NEON AOP Imaging Spectroscopy Survey of Upper East River Colorado Watersheds: Raw-Space Radiance and Observational Variable Dataset. Watershed Function SFA, ESS-DIVE repository. Dataset. doi:el-usecase-1.2 accessed via https://data-dev.ess-dive.lbl.gov/datasets/doi:el-usecase-1.2 on 2022-03-18

Files in this dataset Package: ess-dive-64a126fcda197ec-20210930T145535990704

	Name	File type	Size	Download All
	Metadata: NEON_AOP_Imaging_Spectroscopy_Survey_of_Upper_East_River_Colorado_Watersheds_Raw_Space_Radiance_and_Observational_Variable_Dataset.xml	EML v2.2.0	11 KB	Download

External Links to Data or Metadata

External links for this dataset

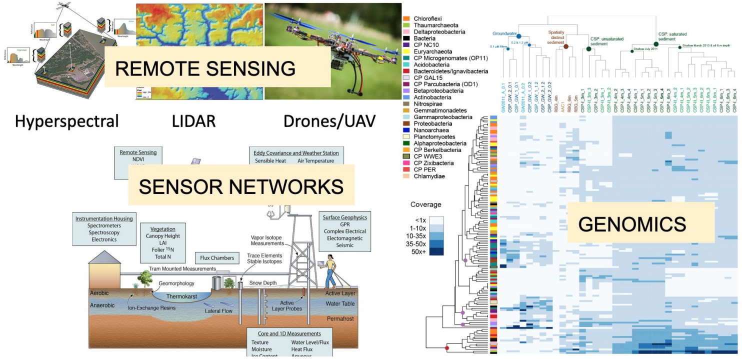
Description	Relationship	URL
Flight Data	[has part] Data that is part of this dataset.	https://portal.nersc.gov/wfsfa/doi-10-15485-1617204/

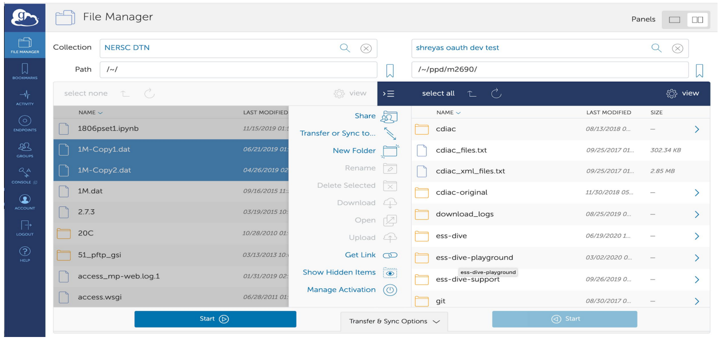
General

Identifier: ess-dive-76578754569b1d4-20210930T145533613560

Supporting Growth In Data

- Increasing need to support large files and datasets
- Enabling multiple ways to make large data uploads seamless
    - ESS-DIVE API
    - Native ESS-DIVE Uploads
    - Globus
- Upcoming: **Tier 2 Storage layer** to for very large datasets
    - Supports very large datasets and files
    - Hierarchical data and folders
    - Direct web UI access to browse and download from Tier 2 storage
    - Globus for High Performance Downloads





Collection NERSC DTN

Path /~/

NAME	LAST MODIFIED
1806pset1.ipynb	11/15/2019 01:
1M-Copy1.dat	06/21/2019 01:
1M-Copy2.dat	04/26/2019 02
1M.dat	09/16/2015 11:
2.7.3	03/19/2015 10:
20C	10/28/2010 01:
51_pftp_gsi	03/13/2013 10:
access_mp-web.log.1	01/31/2019 02:
access.wsgi	06/28/2011 01:

shreyas oauth dev test

/~/ppd/m2690/

NAME	LAST MODIFIED	SIZE
cdiac	08/13/2018 0...	–
cdiac_files.txt	09/25/2017 01...	302.34 KB
cdiac_xml_files.txt	09/25/2017 01...	2.85 MB
cdiac-original	11/30/2018 05...	–
download_logs	08/25/2019 0...	–
ess-dive	06/19/2020 1...	–
ess-dive-playground	03/02/2020 0...	–
ess-dive-support	09/26/2019 0...	–
git	08/30/2017 0...	–

ess-dive-playground

Globus interface for large scale managed transfers

Upcoming: Tier 2 Storage and Large Uploads

ESS-DIVE

Individual Files over 500GB

Datasets containing **any file over 500GB**, such as LIDAR or drone data

Over 100 files outside of Zip file

Datasets containing **over 100 files** that are not stored in a **compressed (or “zipped”) hierarchy** should be treated as large data.

Tier 2 will support uploading large volumes of data; large numbers of files; nested folder structures.

Globus provides high performance interface for data uploads

globus

Tracking Data Versions

- ESS-DIVE tracks all versions of data packages internally *before and after publication*
- Large model data may present challenges with making copies of files with different versions
- Citations indicate version used:

Creator (Publication Year). Title. Publisher. Dataset. Identifier "accessed via data.ess-dive.lbl.gov on YYYY-MM-DD"

Any Questions?

Photo by Towfiqu barbhuiya on Unsplash

Feedback Tell us what you need!

What are your model data archiving needs?

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

ESS Modeling Community Needs Summary

- Data are extremely heterogeneous and increasing in volume and complexity
- Disconnect between model and observational data
- Workflows involving manual retrieval of data are not scalable
- Short-term need for most researchers: guidelines to archive model data - e.g. those associated with publications
- Long-term: Development of a standardized *model-to-archive* pipeline and *data-to-model* pipeline.

Future design based on community needs

- Model-to-archive pipeline
  - Community-informed guidelines on creating standardized model data packages
  - Pathway for model data packages >500 GB size threshold
  - Ability to extract specific subsets of model simulations
  - Project portals for sharing and collaborating on pre-published model data
  - (long-term) Automating the writing and/or organization of files comprising data packages for specific models or journals.
- Data-to-model pipeline
  - Support for data formats typically used in model simulations (e.g., netCDF)
  - Interoperability between individual data packages in ESS-DIVE and other data centers for model-data integration (field observations and remote sensing data to use for model development, parameterization, and performance testing to improve future measurement designs)

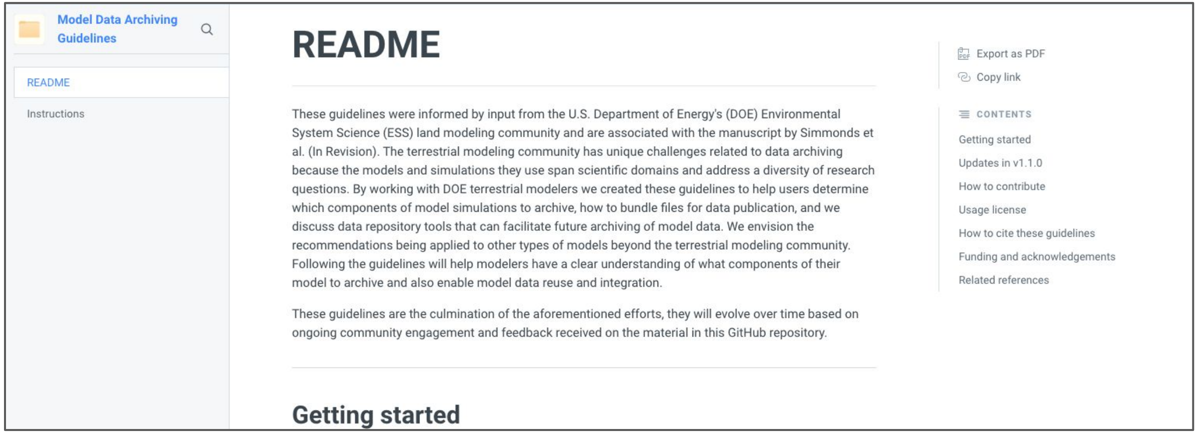
Summary

ESS-DIVE stores model data and has new terrestrial model archiving guidelines

- Model data has many components - inputs, outputs, code, documentation
- Publishing model data is beneficial and enables reuse of simulation data
- Guidelines what to archive and how to split datasets for diverse terrestrial model data used by ESS researchers
- ESS-DIVE has many features supporting model data archiving

Questions? Email ess-dive-support@lbl.gov

You can find all this material on ESS-DIVE's…



README

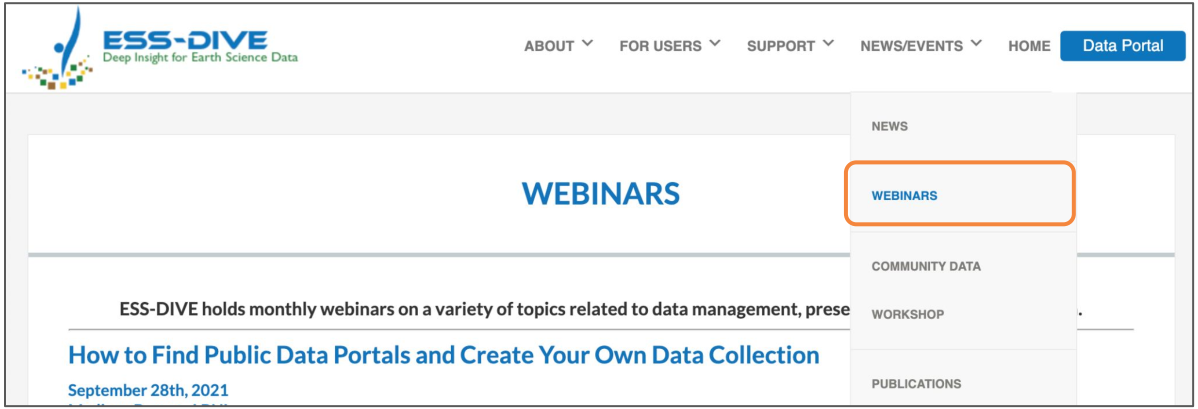
These guidelines were informed by input from the U.S. Department of Energy's (DOE) Environmental System Science (ESS) land modeling community and are associated with the manuscript by Simmonds et al. (In Revision). The terrestrial modeling community has unique challenges related to data archiving because the models and simulations they use span scientific domains and address a diversity of research questions. By working with DOE terrestrial modelers we created these guidelines to help users determine which components of model simulations to archive, how to bundle files for data publication, and we discuss data repository tools that can facilitate future archiving of model data. We envision the recommendations being applied to other types of models beyond the terrestrial modeling community. Following the guidelines will help modelers have a clear understanding of what components of their model to archive and also enable model data reuse and integration.

These guidelines are the culmination of the aforementioned efforts, they will evolve over time based on ongoing community engagement and feedback received on the material in this GitHub repository.

Getting started

Model Data Archiving Documentation page

https://docs.ess-dive.lbl.gov/



WEBINARS

ESS-DIVE holds monthly webinars on a variety of topics related to data management, prese

How to Find Public Data Portals and Create Your Own Data Collection

September 28th, 2021

Webinar page

where this slide deck is available for download

https://ess-dive.lbl.gov/

Additional Resources

Docs: https://ess-dive.gitbook.io/model-data-archiving-guidelines/

Paper: https://datascience.codata.org/articles/10.5334/dsj-2022-003/

Dataset: https://data.ess-dive.lbl.gov/view/doi:10.15485/1813868

Github: https://github.com/ess-dive-community/essdive-model-data-archiving-guidelines

Thank You!

ESS-DIVE

@ESS-DIVE

Join ESS-DIVE’s Community Mailing List!

http://bit.ly/essdiveMailingList

Contact us at ess-dive-support@lbl.gov