Large Data and Model Data Archiving with ESS-DIVE

Fianna O'Brien
Computer Systems Engineer

ESS-DIVE BERKELEY LAB NERSC NCEAS DataONE U.S. DEPARTMENT OF ENERGY Office of Science

2021 Community Data Workshop

Presentation Overview

- What **counts** as “large data”?
- Why is large data **different**?
- How do I **upload** large data?
- How do I **download** large data?
- How do I **organize & document** my large data?

What counts as “large data”?

Large Data Defined

ESS-DIVE

Individual Files over 100GB

Datasets containing **any file over 100GB**, are too large for upload via the web interface or API.

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.

“Download All” wanted for <3GB

Only packages with **<3GB of data can use “Download All”** feature.

If downloading all data at once is **necessary for your users**, your package should be treated as large data.

Trouble Uploading

Even data **files less than 100GB can be difficult to upload** using the API. If you’re having **difficulties uploading**, using the tools for large data may help.

Icons from www.flaticon.com

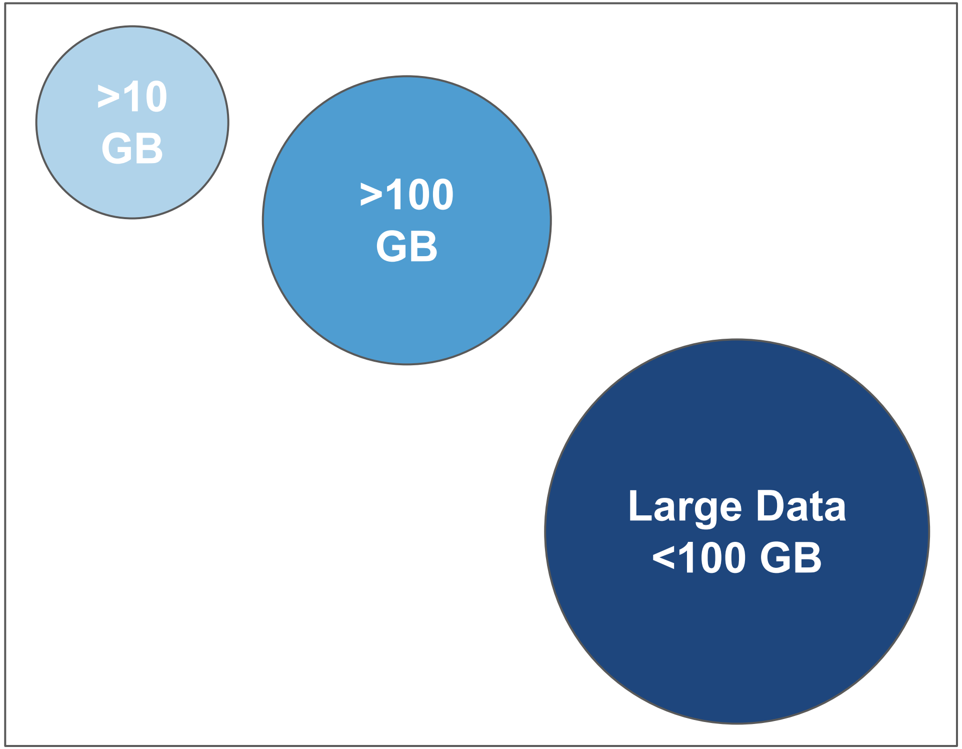
Do I have large data?

If any of the previous definitions fit your data or if you have questions about your specific case, contact us.

Reach out to us at ess-dive-support@lbl.gov

Why is large data different?

Large Data Challenge



Challenge

Upload size limited on web form to <10GB & <100GB via package service.

Solution

Large data can be stored on the ESS-DIVE extended NERSC supercomputer resources.

Large Data Upload Methods

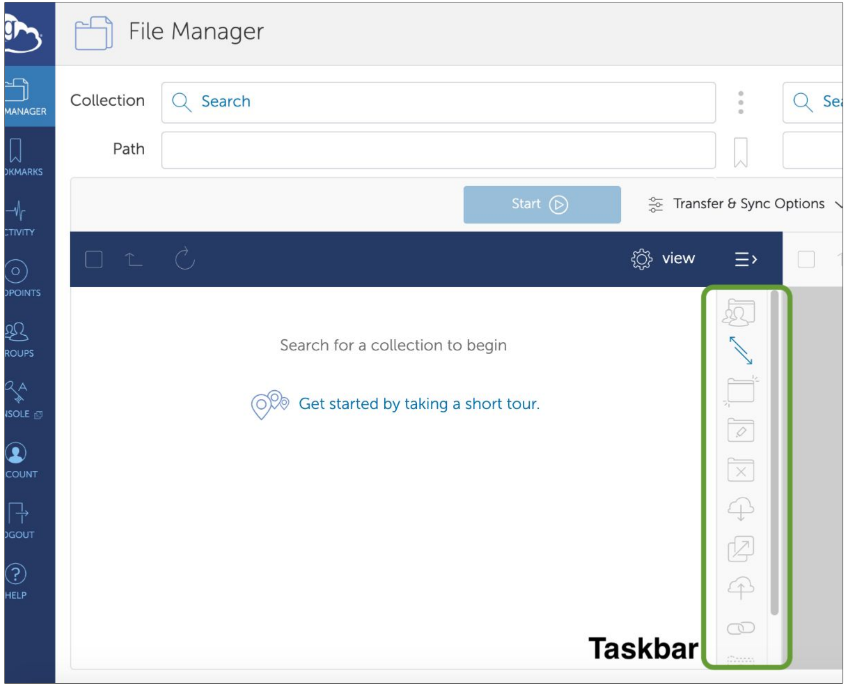
ESS-DIVE

"Download All" uncompressed files	No. of uncompressed files	Volume per file	Upload Method
< 3 GB total	< 100	< 10 GB	Web Upload Form
< 3 GB total	< 100	10-100 GB	Package Service API
> 3 GB total	> 100	> 100 GB	Globus: Data Transfer Service

Uploading large data

What is Globus?

- Free, cloud-based data transfer services for moving significant amounts of data.
- ESS-DIVE uses this to move users' local data to NERSC supercomputer storage



Process Phases

Request Large Data Upload

01

Uploader sends request with description of data to ESS-DIVE support.

ESS-DIVE reviews request and approves uploader for large data.

Create Metadata

02

Uploader creates package metadata and submits for publication.

ESS-DIVE reviews metadata, requests changes, & marks package for Globus upload.

File Upload

03

Uploader uploads data files to Globus via desktop application.

ESS-DIVE confirms transfer & publishes data package with linked NERSC data directory.

Icons from www.flaticon.com

Using Globus for Data Upload

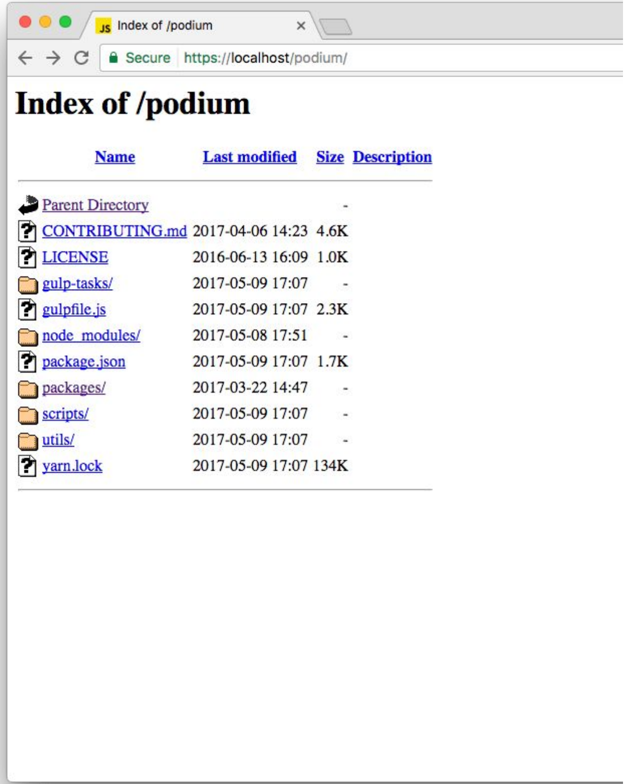
1. Create your account using **ORCID**
2. **Download** desktop application
3. Connect Globus to your **local storage**
4. **Sync data** to ESS-DIVE Tier 2 storage

Demo video available later in the session

Downloading large data

Downloading Large Data from ESS-DIVE

- Link for data access listed in metadata
- Data displayed & downloadable in ESS-DIVE branded **Apache Index**
- **Pros**: Users can explore data hierarchy
- **Cons**: Downloads from Tier 2 are not added to download metrics



Index of /podium

Name	Last modified	Size	Description
Parent Directory		-	
CONTRIBUTING.md	2017-04-06 14:23	4.6K	
LICENSE	2016-06-13 16:09	1.0K	
gulp-tasks/	2017-05-09 17:07	-	
gulpfile.js	2017-05-09 17:07	2.3K	
node_modules/	2017-05-08 17:51	-	
package.json	2017-05-09 17:07	1.7K	
packages/	2017-03-22 14:47	-	
scripts/	2017-05-09 17:07	-	
utils/	2017-05-09 17:07	-	
yarn.lock	2017-05-09 17:07	134K	

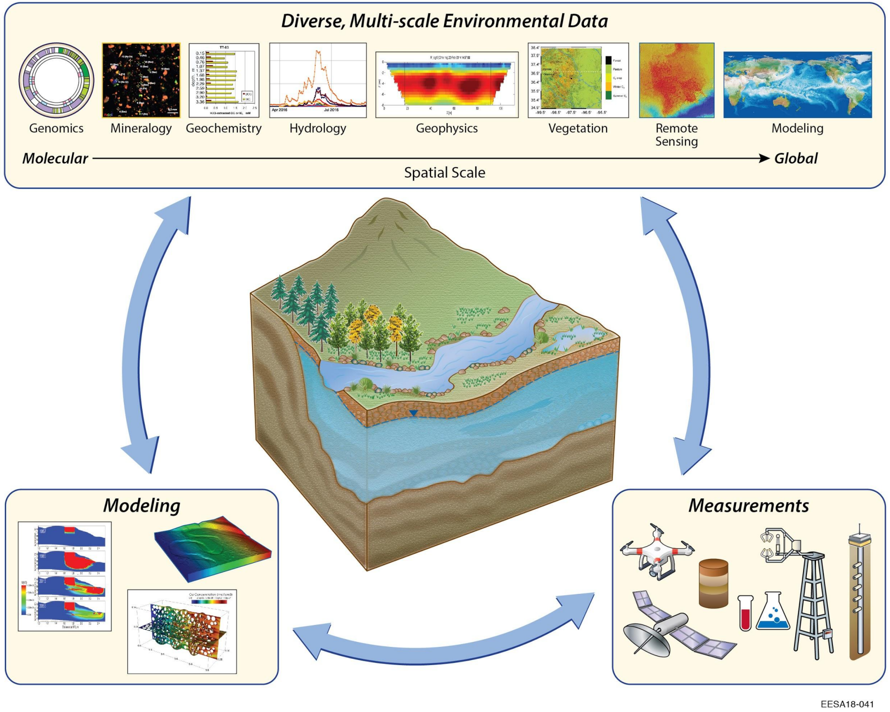
Generic Apache Index Example from Vestride Fancy Index

Organizing & Documenting large data

Research that produces large data

- Model data
- Genomic data
- Vegetation/Remote Sensing

The research topic will inform how to organize research data products into multiple data files and/or into multiple data packages.



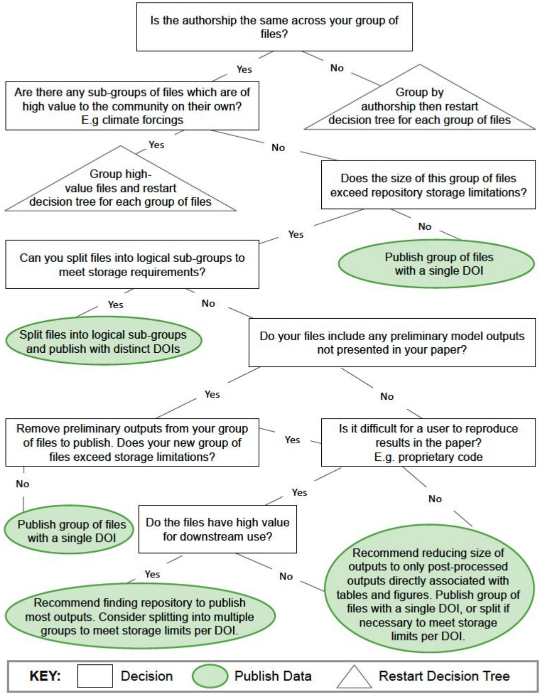
Varadharajan et al. (2019), Launching an accessible archive of environmental data, Eos, 100, https://doi.org/10.1029/2019EO111263.

Decision Tree to Archiving Model Data

ESS-DIVE's research on archiving Model Data recommends separating model data into smaller files according to:

- Authorship
- Downstream value of files
- Repository storage limitations

This decision tree presents the logic used to break down a very large dataset.



Is the authorship the same across your group of files?

Yes

No

Group by authorship then restart decision tree for each group of files

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

Yes

Group high-value files and restart decision tree for each group of files

No

Does the size of this group of files exceed repository storage limitations?

Yes

No

Publish group of files with a single DOI

Can you split files into logical sub-groups to meet storage requirements?

Yes

Split files into logical sub-groups and publish with distinct DOIs

No

Do your files include any preliminary model outputs not presented in your paper?

Yes

Remove preliminary outputs from your group of files to publish. Does your new group of files exceed storage limitations?

No

Publish group of files with a single DOI

Yes

No

Is it difficult for a user to reproduce results in the paper? E.g. proprietary code

Yes

Do the files have high value for downstream use?

Yes

Recommend finding repository to publish most outputs. Consider splitting into multiple groups to meet storage limits per DOI.

No

No

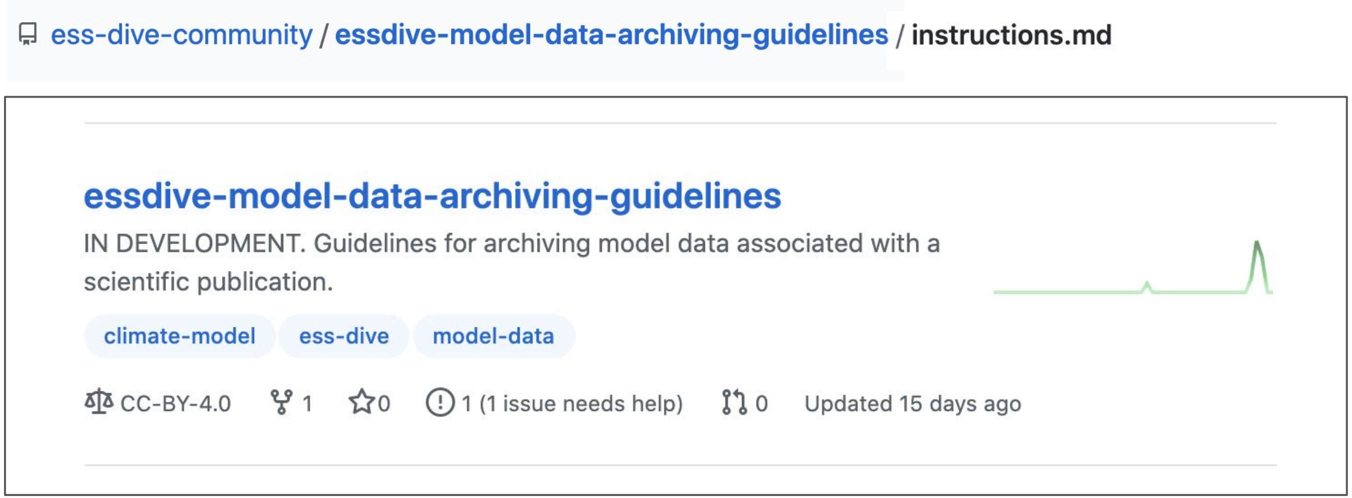
Recommend reducing size of outputs to only post-processed outputs directly associated with tables and figures. Publish group of files with a single DOI, or split if necessary to meet storage limits per DOI.

KEY: □ Decision ○ Publish Data △ Restart Decision Tree

"Deciding how to bundle files" from ESS-DIVE Model Data Archiving GitHub

Model Data Archiving Tutorial

- 1:00 - 2:00 pm PST during Reporting Format breakouts
- How to use available Model Data Archiving instructions to start organizing your model data today! Intended to be applied to both large and small model data volumes.



ess-dive-community / **essdive-model-data-archiving-guidelines** / **instructions.md**

essdive-model-data-archiving-guidelines

IN DEVELOPMENT. Guidelines for archiving model data associated with a scientific publication.

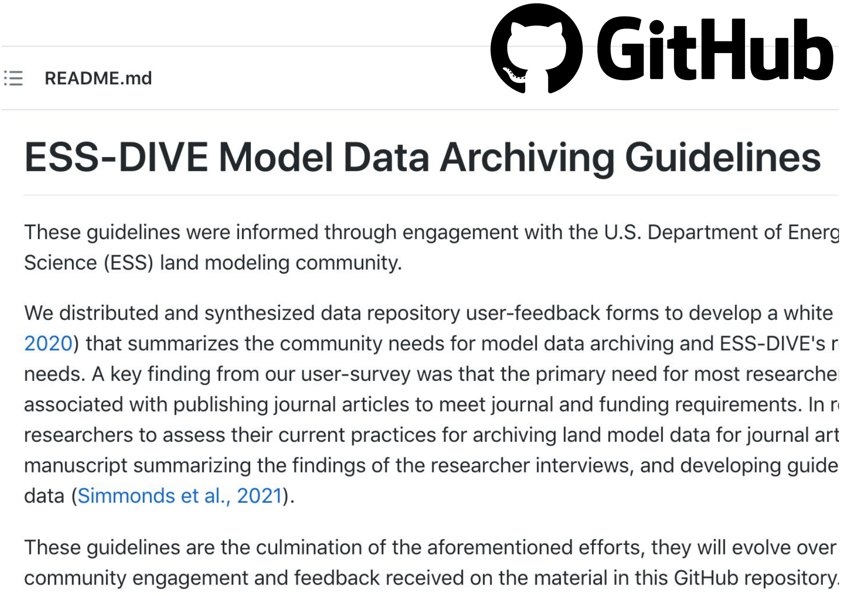
climate-model ess-dive model-data

CC-BY-4.0 1 0 1 (1 issue needs help) 0 Updated 15 days ago

Model Data Archiving Guidelines

ESS-DIVE has created the Model Data Archiving Guidelines GitHub, which includes decision trees for:

- Choosing files to include
  - Recommendations for important details and file formatting
- Deciding how to bundle files
  - Considering authorship, storage limitations, and downstream value
- File Level Metadata Guidelines
  - Guidance for creating standardized data dictionary & file catalogs



ESS-DIVE Model Data Archiving Guidelines

These guidelines were informed through engagement with the U.S. Department of Energ Science (ESS) land modeling community.

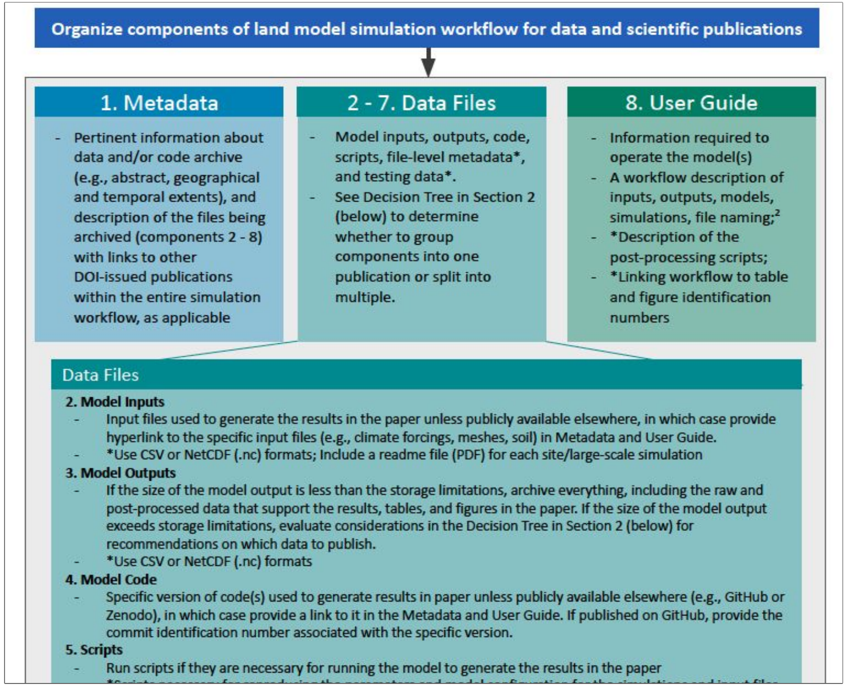
We distributed and synthesized data repository user-feedback forms to develop a white 2020) that summarizes the community needs for model data archiving and ESS-DIVE's r needs. A key finding from our user-survey was that the primary need for most researche associated with publishing journal articles to meet journal and funding requirements. In r researchers to assess their current practices for archiving land model data for journal art manuscript summarizing the findings of the researcher interviews, and developing guide data (Simmonds et al., 2021).

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

"Files to Include" from ESS-DIVE Model Data Archiving GitHub

Files to Include in Model Data Package

- Metadata
  - Describes data files & provides information about data/code
- Data Files
  - Model Inputs & Outputs
  - Model Code
  - Scripts
  - File Level Metadata (optional)
  - Model Testing Data (optional)
- User Guide
  - Guide for running model, workflow of inputs & outputs.



Organize components of land model simulation workflow for data and scientific publications

1. Metadata

- Pertinent information about data and/or code archive (e.g., abstract, geographical and temporal extents), and description of the files being archived (components 2 - 8) with links to other DOI-issued publications within the entire simulation workflow, as applicable

2 - 7. Data Files

- Model inputs, outputs, code, scripts, file-level metadata*, and testing data*.
- See Decision Tree in Section 2 (below) to determine whether to group components into one publication or split into multiple.

8. User Guide

- Information required to operate the model(s)
- A workflow description of inputs, outputs, models, simulations, file naming;[2]
- *Description of the post-processing scripts;
- *Linking workflow to table and figure identification numbers

Data Files

2. Model Inputs

- Input files used to generate the results in the paper unless publicly available elsewhere, in which case provide hyperlink to the specific input files (e.g., climate forcings, meshes, soil) in Metadata and User Guide.
- *Use CSV or NetCDF (.nc) formats; Include a readme file (PDF) for each site/large-scale simulation

3. Model Outputs

- If the size of the model output is less than the storage limitations, archive everything, including the raw and post-processed data that support the results, tables, and figures in the paper. If the size of the model output exceeds storage limitations, evaluate considerations in the Decision Tree in Section 2 (below) for recommendations on which data to publish.
- *Use CSV or NetCDF (.nc) formats

4. Model Code

- Specific version of code(s) used to generate results in paper unless publicly available elsewhere (e.g., GitHub or Zenodo), in which case provide a link to it in the Metadata and User Guide. If published on GitHub, provide the commit identification number associated with the specific version.

5. Scripts

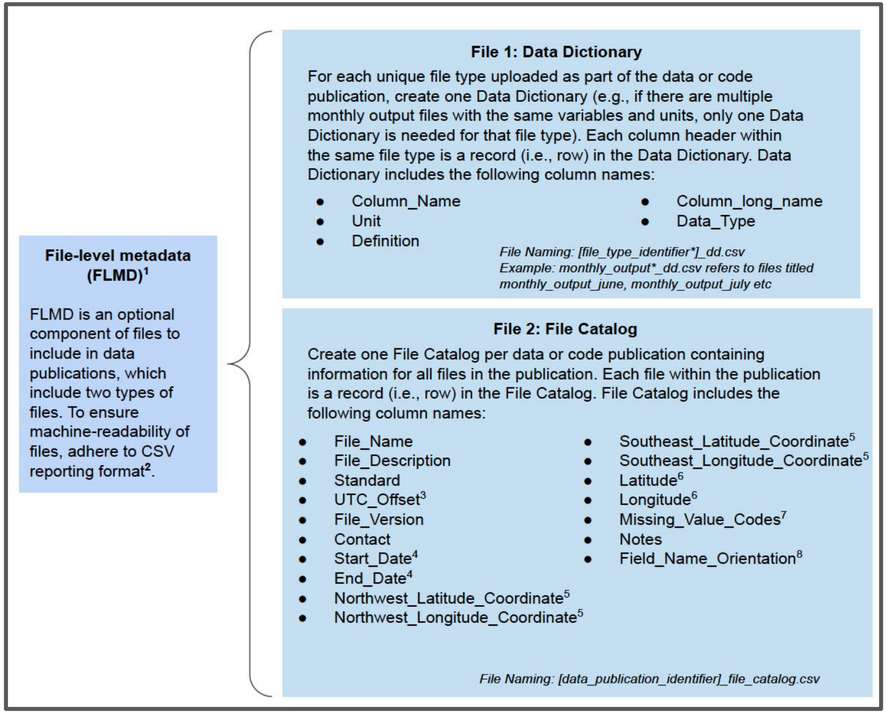
- Run scripts if they are necessary for running the model to generate the results in the paper

"Files to Include" from [ESS-DIVE Model Data Archiving GitHub]

File Level Metadata

A reporting format that can be applied to model data archiving

- Data Dictionary
  - Explains data file column headers, including variable full name, description, units, and data type
- File Catalog
  - Documents files contained in data package, including file name, description, version, as well as data collection location, dates, and data standard.



File-level metadata (FLMD)[1]

FLMD is an optional component of files to include in data publications, which include two types of files. To ensure machine-readability of files, adhere to CSV reporting format[2].

File 1: Data Dictionary

For each unique file type uploaded as part of the data or code publication, create one Data Dictionary (e.g., if there are multiple monthly output files with the same variables and units, only one Data Dictionary is needed for that file type). Each column header within the same file type is a record (i.e., row) in the Data Dictionary. Data Dictionary includes the following column names:

- Column_Name
- Unit
- Definition
- Column_long_name
- Data_Type

File Naming: [file_type_identifier]_dd.csv*
*Example: monthly_output*_dd.csv refers to files titled monthly_output_june, monthly_output_july etc*

File 2: File Catalog

Create one File Catalog per data or code publication containing information for all files in the publication. Each file within the publication is a record (i.e., row) in the File Catalog. File Catalog includes the following column names:

- File_Name
- File_Description
- Standard
- UTC_Offset[3]
- File_Version
- Contact
- Start_Date[4]
- End_Date[4]
- Northwest_Latitude_Coordinate[5]
- Northwest_Longitude_Coordinate[5]
- Southeast_Latitude_Coordinate[5]
- Southeast_Longitude_Coordinate[5]
- Latitude[6]
- Longitude[6]
- Missing_Value_Codes[7]
- Notes
- Field_Name_Orientation[8]

File Naming: [data_publication_identifier]_file_catalog.csv

"Files Level Metadata" from *ESS-DIVE Model Data Archiving GitHub*

Polls

Poll questions

- Do you plan to upload any data to ESS-DIVE?
  - ○ Yes, large data
  - ○ Yes, but not large data
  - ○ Yes, but not sure
  - ○ I have already uploaded to ESSDIVE

Poll questions

- Do you think your data qualifies as Large Data?
    - ○ Yes, I’m certain my data qualifies as Large Data
    - ○ No, my data is not Large Data
    - ○ I’m not sure if my data qualifies as large data

Poll questions

- What kind of data do you plan to submit to ESS-DIVE
  - OPEN ENDED

Poll questions

- Do you have experience with Globus?
    - ○ Yes
    - ○ No

UAS/Spatial Data Archiving

Open Discussion

Conversation Starters

- What have been the difficulties you’ve found sharing large data?
- What have been the most successful experiences sharing large data?
- What have been your experiences in accessing & downloading large data?
    - Using an archive’s main repository?
    - Outside of main repository?