

# Curating simulation-based research

Data Curation Network Lesson Plan





# Meet your instructors

L. Wynholds, University California, Los Angeles

Heather Shimon, University of Wisconsin, Madison

Wanda Marsolek, University of Minnesota, Twin Cities

Fernando Rios, University of Arizona

Girmaye Misgna, University of Pennsylvania



# Outline

- Introduction & goals
- Overview of simulation data
- CURATED Steps Discussion
  - Check
  - Understand
  - Request
  - Augment
  - Transform
  - (Document)
  - (Evaluate)

# Introduction

## Goals

- Understand what simulation-based research is
- How to recognize it
- Know what kinds of questions to ask of researchers
- Understand related curation resources

## Scope

- Simulation models, not AI/ML, not animal models
- Assume you have already reviewed the CURATED model in general
- Assume you've taken (or are taking) the Code workshop

# Activity 1

<https://www.menti.com/XXXXX>

Menti.com code: XXXX XXXX

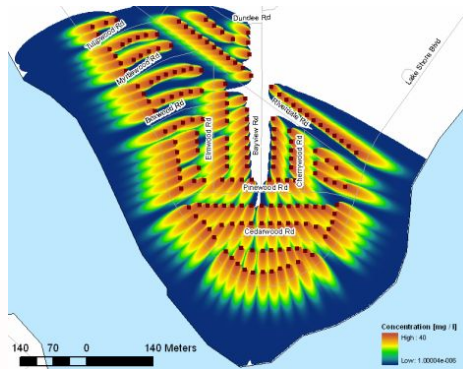
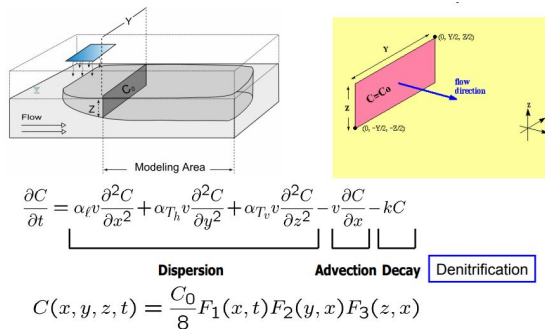


# What is simulation?

Conceptual model -> mathematical model -> computer code -> output



USGS Scientific Investigations Report 2008–5220



```

120 metadata_tags = Metadata(self.config, metadata_path,
121                          self.log).parse_metadata()
122
123 if not metadata_tags:
124     return Status.INVALID_CONFIG
125
126 if self.wasabi_dart_hostbucket override:
127     with open(self.workflow, "r") as f:
128         wkfl_json = json.load(f)
129         if not 'storageServices' in wkfl_json:
130             print('storageServices key not found in DART workflow file')
131             return Status.INVALID_CONFIG
132         for item in wkfl_json['storageServices']:
133             item['host'] = self.wasabi_s3host
134             item['bucket'] = self.wasabi_s3bucket
135             self.workflow_file = NamedTemporaryFile(prefix="rebach", mode="w", delete=True)
136             self.workflow_file.write(json.dumps(wkfl_json))
137             self.workflow_file.flush()
138             self.workflow = self.workflow_file.name
139
140 return bag_name, metadata_tags

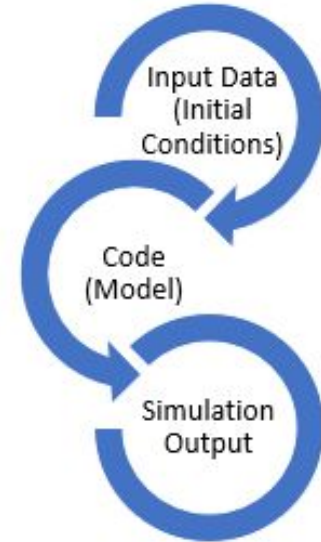
```

# Simulation-based research

## Examples:

- Groundwater flow simulation
- Black hole models
- Simulation of the effects of drugs on cells
- Financial simulations
- Climate simulation
- Molecular dynamics simulations
- Plate tectonics
- ...and many more

## Generalized Overview of Simulation Data Processes (for Curation Purposes):



Simulation data and models

# Simulation vs AI/ML

## Simulation

- Given inputs, predicts outputs
- Models a real-world phenomenon using a **defined mathematical model**
- Small amount of additional data needed for model calibration
- Can be used to simulate systems if the model's underlying assumptions are met.
  - e.g., groundwater model: easy to change the soil properties

## AI/ML

- Given inputs, predicts outputs
  - But can also do other things (classify)
- Models a real-world phenomenon where the **model is a black box**
  - Driven by the training process
- Needs lots of data
- Risks applying the model to cases that differ from the training data
  - E.g., groundwater model: a change in soil properties may require re-training.



# Common terms you might hear

[See the glossary  
handout](#)

**Model:** A simplified representation of reality. Usually a model is expressed mathematically so it can be simulated in a computer.

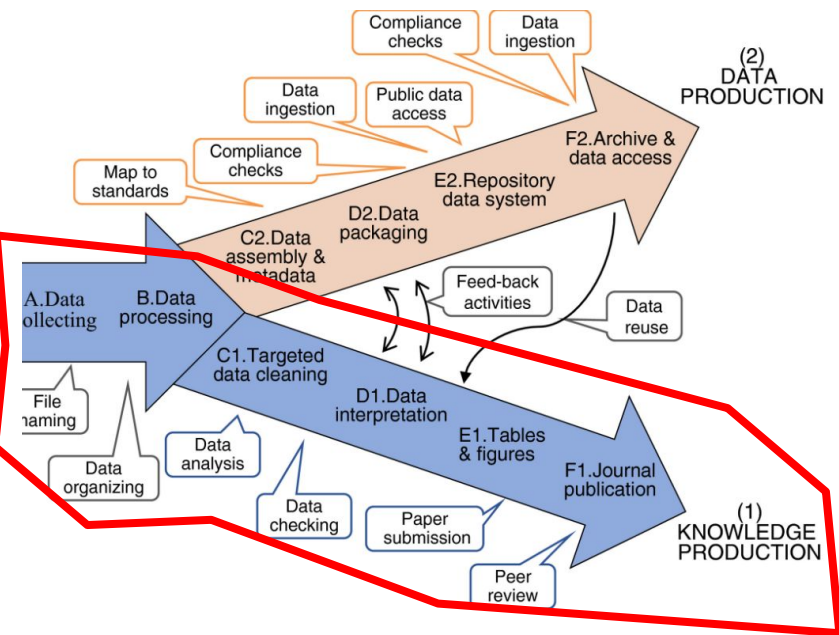
**Simulation:** Computer approximation of reality, based on a model

**Parameters, forcing factors, boundary conditions, initial conditions:** Tunable 'dials' of a simulation that cause it to behave differently

**Experiment:** A particular simulation run. A set of runs with differing parameters or model characteristics is sometimes called an ensemble

**Pipeline, workflow:** a set of usually sequential computational steps that produces an output. Add simulation = Simulation-based research

# Characteristics of simulation data



Baker & Mayernik (2020). CC By 4.0  
<https://doi.org/10.1002/ecs2.3191>

Mullendore et al. (2021) expand on Baker & Mayernik's (2020) conceptual model

Two streams with different characteristics

- Data production
  - Producers of reference data (e.g., gov't agencies)
  - Requires well-planned and funded data curation support
- Knowledge production
  - Data re-users
  - Extract knowledge from data
  - Can produce knowledge despite minimal curation

Focus on knowledge production in this workshop

- Most curation @ universities is for KP

# What is special about simulation data?

Data itself, not much, really

- Isn't really a data format
- Refer to the other DCN primers

For simulation processes on the other hand...

- Holistic view towards improving reusability
- Advice for researchers?
- Structuring simulation materials (data, code)
- Ethical considerations, e.g., computational requirements
- What to keep? [EarthCube Rubric](#)



What About Model Data? Best Practices for Preservation and Replicability

Theme: Community Commitment

Is it anticipated that your simulation workflow outputs will have broad community impact and downstream reuse?

**Question 1**

Used in a "Highly Influential Scientific Assessment"

*As defined, for example, by OMB "Revised Information Quality Bulletin for Peer Review" (2004 Apr 15): a scientific assessment whose "dissemination could have a clear and substantial impact on important public policies (including regulatory actions) or private sector decisions with a potential effect of more than \$500 million in any one year or that the dissemination involves precedent setting, novel and complex approaches, or significant interagency interest."*

- ☒ No, not used in any HISA.
- ☐ Subset of output may enable fact checking, e.g. all output are not needed, but selected or derived products (e.g. ensemble mean and spread) will provide adequate scientific representation.
- ☐ Used in a HISA. Need to keep output for future fact checking.

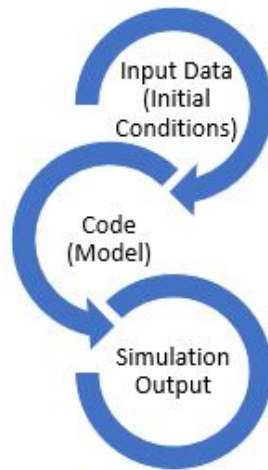
**Question 2**

# Common elements of simulation models

## Pieces of models and related primers

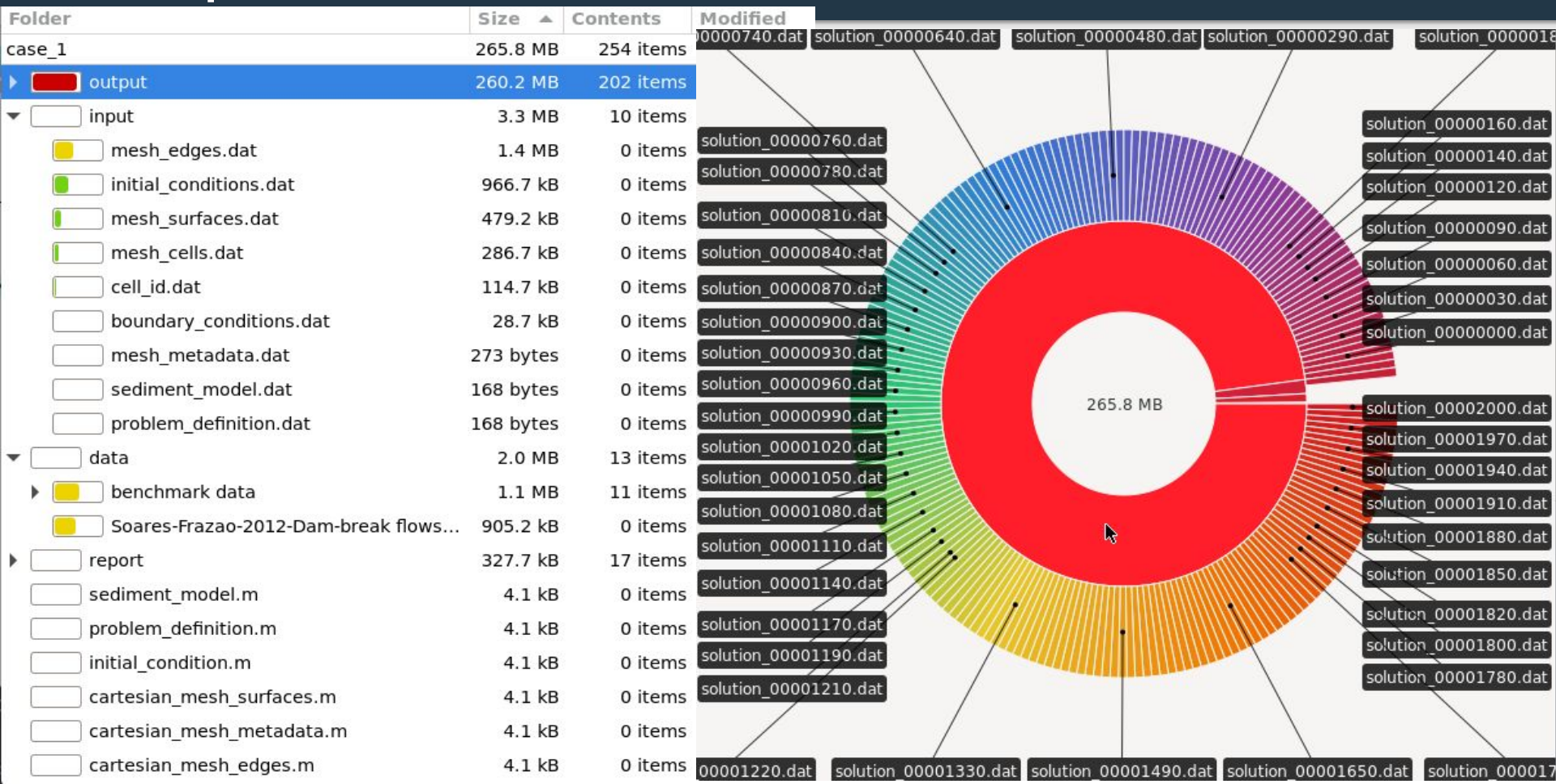
- Input files
  - Could include many kinds/formats
  - Initial conditions, calibration files, parameter values
- One or more software artifacts
  - May be included in the dataset or external/3rd party
  - May or may not have source code
  - Order of execution important
  - May have tight or loose couplings to other models
- Outputs
  - Ideal, but not always present
- Documentation

## Generalized Overview of Simulation Data Processes (for Curation Purposes):



More detail provided as we go through the curated steps

# Example



# CURATE(D) steps and Simulations

What is the goal of curation (for knowledge production)? Mullendore et al. (2021)

- Future understanding and knowledge production, not bitwise reproducibility (can vary by discipline)
- Less focus on raw data, more on driving factors and derived information
- Competing interests (e.g., storage limitations) might change the calculus


## CURATE(D) Steps

- Steps are explanatory
  - Highly non-linear in reality
- Will be calling out to many other primers
  - Simulation models/data aren't that special
- Simulation-specific tips will be sprinkled throughout



# (D) Curator Log

## Curator Log Template

- We can use the curator log throughout the curation process, curation is non-linear
  - Important to document questions, concerns, missing pieces of the puzzle throughout the process and not just at the end. 
  - Lots of moving parts, competing licenses, adequate detail is necessary
- Worksheet for evaluating metadata quality

---

```
Curation log for: <title of dataset>
Handle:
Corresponding researcher name and email:
Curator:
Metadata log created:
```

```
*****
Files received:
*****
```

```
*****
Changes made to files:
*****
```

```
*****
Metadata Changes
*****
```

```
*****
Correspondence Notes
*****
```

```
*****
Other issues
*****
```

```
*****
Original Metadata from Author:
*****
```

# (E) Evaluate

Throughout the process & iterative



- Goal 1: Make the dataset more FAIR
- Goal 2: Review data for ethical concerns resourcing CARE and FATE

Focus on ethics regarding

- Code
- Model
- Initial data
- Human participants

Worksheet for evaluating metadata quality

Metadata Checklist:	
Data Type	<ol style="list-style-type: none"><li>What is the data type, format, size, and number of files?</li><li>Whether the dataset is a subset, part, or related to other datasets</li><li>Does the data description include the level of aggregation, de-identification, processing, or cleaning that was done in the data?</li><li>Does the metadata include the source(s) of any secondary data or previously collected data used in the production of the dataset (e.g. inputs, calibration, reference datasets, ground truthing datasets, etc)</li><li>Metadata available that describes the above adequately for others to use (e.g. readme.txt)</li></ol>
Related Tools, Software, Code & Publications	<ol style="list-style-type: none"><li>State whether specialized tools are needed</li><li>State which code versions and which OS were used to produce the data</li><li>State how to access the code/software/tools (e.g. citation to code or download website)</li><li>State whether the dataset was used in articles or other publications, and if so, citations to published or in-process title/author/journal</li><li>Reference code group mat'ls (link)</li></ol>
Data Standards	<p>Are there data standards used in the dataset? If so, which ones and where are they specified? (cite which standard)</p> <ol style="list-style-type: none"><li>Metadata schemas</li><li>Standard terminologies/controlled vocabularies</li><li>Content/encoding standards</li><li>Common data elements</li><li>Identifiers (PIDs)</li></ol>
Data Access	<ol style="list-style-type: none"><li>Is the DOI and/or other PID listed in the metadata?</li><li>Are there any legal, technical or ethical considerations with sharing the data?</li><li>Did the funders require that the data be made accessible to the public? If so is that information provided to the end users with consideration for future use?</li><li>Did the funders require that the data be available for a set period? If so, how long/what dates?</li></ol>
Data Restrictions	<ol style="list-style-type: none"><li>Are there limitations to use? (e.g. IP restrictions, embargos, requires human subject IRB review for use, etc)</li><li>Are there instructions for gaining access to restricted datasets?</li><li>Embargo period?</li><li>Other access controls?</li></ol>
Data Stewardship	<ol style="list-style-type: none"><li>Who was responsible for producing/QC for the data?</li></ol>
Data Provenance	<ol style="list-style-type: none"><li>Who can be contacted with questions regarding the data? (usually the first</li></ol>



# Concrete dataset examples

## Dataset A

Yang, Huang; Waugh, Darryn W., 2020, "Data associated with Yang et al., 2020, Dependence of Atmospheric Transport into the Arctic on Extent of the Hadley Cell", <https://doi.org/10.7281/T1/AWJUGZ>, Johns Hopkins Research Data Repository, V1

Data associated with Yang et al., 2020, Dependence of Atmospheric Transport into the Arctic on Extent of the Hadley Cell

Version 1.1



Yang, Huang; Waugh, Darryn W., 2020, "Data associated with Yang et al., 2020, Dependence of Atmospheric Transport into the Arctic on Extent of the Hadley Cell", <https://doi.org/10.7281/T1/AWJUGZ>, Johns Hopkins Research Data Repository, V1

[Cite Dataset](#) ▼

[Learn about Data Citation Standards.](#)

[Access Dataset](#) ▼

[Contact Owner](#)

[Share](#)

Dataset Metrics ?

162 Downloads ?

Description ?

Model output (annual-mean values for 6 years) from the GFDL dry dynamical core, see details in Yang et al. (2020), Geophysical Research Letters. (2019-05)

Subject ?

Earth and Environmental Sciences

Keyword ?

atmospheric transport, Eulerian mean circulation, Arctic atmospheric composition

# Concrete dataset examples

## Dataset B

Do, H. X., Smith, J. P., Fry, L. M., Gronewold, A. D. (2020). Monthly water balance estimates for the Laurentian Great Lakes from 1950 to 2019 (v1.1) [Dataset], University of Michigan - Deep Blue Data.

<https://doi.org/10.7302/tx97-nn12>

Title: Monthly water balance estimates for the Laurentian Great Lakes from 1950 to 2019 (v1.1) [Open Access](#) [Deposited](#)

Attribute	Value
Methodology	The new estimates of the water balance of the Laurentian Great Lakes were generated using the Large Lakes Statistical Water Balance Model (L2SWBM). The L2SWBM used multiple independent data sets to obtain the prior distributions and likelihood functions, which were then assimilated by a Bayesian fra... <a href="#">[more]</a>
Description	This data set contains a new monthly estimate of the water balance of the Laurentian Great Lakes, the largest freshwater system on Earth, from 1950 to 2019. The source codes and inputs to derive the new estimates are also included in this dataset.
Creator	<a href="#">Do, Hong X.</a> <a href="#">Smith, Joeseeph P.</a> <a href="#">Fry, Lauren M.</a> <a href="#">Gronewold, Andrew D.</a>
Depositor	hongdo@umich.edu

# Concrete dataset examples

## Dataset C

Horna Munoz, Daniel; Constantinescu, George; Rhoads, Bruce ; Lewis, Quinn; Sukhodolov, Alexander (2020): Confluence Density Effects Simulation Database. University of Illinois at Urbana-Champaign. [https://doi.org/10.13012/B2IDB-6257171\\_V1](https://doi.org/10.13012/B2IDB-6257171_V1)

## Confluence Density Effects Simulation Database

### Citation:

Horna Munoz, Daniel; Constantinescu, George; Rhoads, Bruce ; Lewis, Quinn; Sukhodolov, Alexander (2020): Confluence Density Effects Simulation Database. University of Illinois at Urbana-Champaign. [https://doi.org/10.13012/B2IDB-6257171\\_V1](https://doi.org/10.13012/B2IDB-6257171_V1) [Export Citation](#)

### Persistent link for this dataset:

[https://doi.org/10.13012/B2IDB-6257171\\_V1](https://doi.org/10.13012/B2IDB-6257171_V1)

### Dataset Description

This data set shows how density effects have an important influence on mixing at a small river confluence. The data consist of results of simulations using a detached eddy simulation model.

### Subject

[Physical Sciences](#)

### Keywords

confluence; flow dynamics; density effects

### License

CC0

### Funder

U.S. National Science Foundation (NSF) - **Grant:** BCS 1359911

### Corresponding Creator

Daniel Horna Munoz

# Activity 2: What makes this simulation data?

Guided exercise using sample dataset A


<https://doi.org/10.7281/T1/AWJUGZ>

To get started

- Look over the record page
- View the README
- View the files

Data associated with Yang et al., 2020, Dependence of Atmospheric Transport into the Arctic on Extent of the Hadley Cell

Version 1.1



Yang, Huang; Waugh, Darryn W., 2020, "Data associated with Yang et al., 2020, Dependence of Atmospheric Transport into the Arctic on Extent of the Hadley Cell", <https://doi.org/10.7281/T1/AWJUGZ>, Johns Hopkins Research Data Repository, V1

Cite Dataset ▾ Learn about Data Citation Standards.

Access Dataset ▾

Contact Owner Share

Dataset Metrics ⓘ  
162 Downloads ⓘ

Description ⓘ  
Model output (annual-mean values for 6 years) from the GFDL dry dynamical core, see details in Yang et al. (2020), Geophysical Research Letters. (2019-05)

Subject ⓘ  
Earth and Environmental Sciences

Keyword ⓘ  
atmospheric transport, Eulerian mean circulation, Arctic atmospheric composition

- Any **keywords** that stand out?
- Do the **data files** provide any clues?
- Any **software** clues?
- Do the necessary **files** appear to be there?

# Activity 2: What makes this simulation data?

This dataset contains the GFDL dry model output that are used in the study of “Dependence of Atmospheric Transport into the Arctic on Extent of the Hadley Cell”.

How to cite the data:

Yang, Huang; Waugh, Darryn W., 2019, “Data associated with Yang et al., 2019, Dependence of Atmospheric Transport into the Arctic on Extent of the Hadley Cell”, <https://doi.org/10.7281/T1/AWJUGZ>, Johns Hopkins University Data Archive.

There are 7 subdirectories inside corresponding to the 7 experiments with varied locations of midlatitude jet and the Hadley Cell (HC) extent following the method of Garfinkel et al. (2013).

The name of these experiments are as “A?”, in which ? denotes the difference in parameter A (see al. (2019)) among the experiments that act to perturb the location of jet and the HC extent. We vary A with an even interval of 5. When A = 0, the corresponding experiment “A0” (a.k.a. control run) radiative forcing as the classic Held and Suarez (1994) benchmark.

In each subdirectory (i.e., each control or perturbed experiment), there are 4 types of netcdf files (for Matlab) restoring the annual-mean output.

Differences among the 4 types of netcdf files are marked by the tag “atmos\_\*”, where \* can be:

- (a) daily: files with meteorological fields (annual average from daily instantaneous fields);
- (b) dailym: files with tracer fields (annual average from daily average fields) in which grid30L50asia30L50 is the ZA tracer analyzed in Yang et al. (2019);
- (c) eddy: files with eddy flux;
- (d) mean: files with zonal-mean flux.

Files initiated with numbers are the annual-mean output for each year, while those initiated with concatenated 6-year (avoid first 4 years as spin-up) output.

The other 3 mat files are:

- (a) HClat\_monthly.mat: diagnosed HC extent;
- (b) Jetlat.mat: diagnosed jet latitude;
- (c) theta\_daily.mat: diagnosed isentropic surfaces.

If you have any further question associated with the data, please do not hesitate to reach the corresponding author, Huang Yang at [hyang19@atmos.ucla.edu](mailto:hyang19@atmos.ucla.edu).

- Any **keywords** that stand out?
- Do the **data files** provide any clues?
- Any **software** clues?
- Do the necessary **files** appear to be there?

# (C) Check

The 'CURATE(D)' CHECK step revolves around inventorying and reviewing the contents of the dataset and verifying that it is appropriate for the repository.

This often includes:

- Review to ensure data is in scope for the repository
- Inventory the contents of the data files (e.g., open and sample the files or code)
- Verify all metadata provided by the researcher; check available documentation

## (C) Check cont'd

Simulation data tends to be the result of interactions between models, datasets and code. The check step needs to document how all of the interactions, data, and workflows were produced, including documenting any code used to produce the dataset.

When checking simulation data pay special attention to checking for:

- Code, software, and/or computing environments used
- Related datasets, materials, documentation or publications
- Documentation of data production
- References to data formats/metadata standards/code used/programming languages used

# Checking for Ethical Concerns (Check/Evaluate)

Evaluating data for potential issues:

Where to look:

- Code
- Model
- Initial conditions/input data/sources

What to look for:

- Human Subjects
- Sensitive Species
- Politically contentious topic areas
- Licensing Agreements
- Accessibility

Ethics of access to publications (e.g. publication being behind a paywall)



# Group Discussion

With your group/table discuss

- What kinds of ethical issues might crop up in your discipline?
- What might YOU look for when it comes to these issues?

# (U) Understand

- After inventorying and reviewing the contents, the **Understand** step is a closer examination of:
  - what they are
  - how they interrelate
  - what information is needed for reuse
- **Check** and **Understand** steps generate the information needed for the **Request** step.

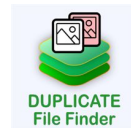
# (U) Understand step often includes

- **Examining** for quality assurance and usability issues
- **Assessing** for ethical issues including sensitive data, consent, risk to persons/places, licensing/permissions, accessibility, etc. (see “Checking for Ethical Concerns” slide)
- **Determining** if the documentation is sufficient for peer researchers to understand and reuse the contents
- **Understanding** how the files relate to each other

# (U) Understand step: Simulations specifically

## Focus on how the files relate to each other

- Simulation data can consist of sequential files - check for missing files in a sequence
- Often, experiments are repeated and datasets will contain multiple similar folders and files
  - Check that naming conventions are consistent
  - Note that a standard duplicate file check will usually reveal multiple identical files – for simulation data, this is often ok, since only small things may change between multiple runs of the model



- Make sure it's clear which outputs correspond to which inputs
- If available, review the related publication for methods

# (U) Understand: DCN primers for data formats

- Simulation data can be in many formats
- **Data curation primers** useful when working with simulation data
  - Accessibility Data Primer
  - Acrobat PDF Primer
  - Column Binary Data Curation Primer
  - Jupyter Notebooks Primer
  - Matlab Primer
  - Microsoft Access Primer
  - Microsoft Excel Primer
  - netCDF Primer
  - R Primer

# (U) Understand step: Which parts are needed?

## Knowledge production vs data production

- “... primary goal of most projects involving computer simulations is to **increase scientific knowledge** and the simulations are used as a tool to that end.”
- “... we are producing far more simulation output than can be reasonably stored in repositories. **Knowledge production research should preserve minimal output in repositories.**” Mullendore et al. (2021)

## Components to preserve and share (this varies)

- **input data**: initial conditions, calibration files, parameter values
- **code/model**: preprocessing, configuration, post processing
- **outputs**: depends on the simulation
- **documentation**

# What do these things look like?

## Example Dataset B

Do, H. X., Smith, J. P., Fry, L. M., Gronewold, A. D. (2020). Monthly water balance estimates for the Laurentian Great Lakes from 1950 to 2019 (v1.1) [Dataset], University of Michigan - Deep Blue Data.

<https://doi.org/10.7302/tx97-nn12>

Title: Monthly water balance estimates for the Laurentian Great Lakes from 1950 to 2019 (v1.1)

[Open Access](#) [Deposited](#)

Attribute	Value
Methodology	The new estimates of the water balance of the Laurentian Great Lakes were generated using the Large Lakes Statistical Water Balance Model (L2SWBM). The L2SWBM used multiple independent data sets to obtain the prior distributions and likelihood functions, which were then assimilated by a Bayesian fra... <a href="#">[more]</a>
Description	This data set contains a new monthly estimate of the water balance of the Laurentian Great Lakes, the largest freshwater system on Earth, from 1950 to 2019. The source codes and inputs to derive the new estimates are also included in this dataset.
Creator	<a href="#">Do, Hong X.</a> <a href="#">Smith, Joeseeph P.</a> <a href="#">Fry, Lauren M.</a> <a href="#">Gronewold, Andrew D.</a>
Depositor	hongdo@umich.edu

# Example Dataset B







## Review list of files

- Documentation, inputs, model/source code, and outputs
- README provides overview, software specification, variable definitions
  - *"This data set contains new estimates of the Great Lakes water balance together with the L2SWBM source code and inputs synthesized for this project..."*
- Source code R files are organized in run order with config\_README

[https://deepblue.lib.umich.edu/data/concern/data\\_sets/sb3978457](https://deepblue.lib.umich.edu/data/concern/data_sets/sb3978457)

Files [Readme](#) [Provenance Log](#)

Files (Count: 7; Size: 4.41 MB)

Title	Original Upload	Last Modified	File Size	Access	Actions
 <a href="#">GreatLakesWaterBalanceData_19502...e.txt</a>	2020-06-05	2020-06-05	7.46 KB	<a href="#">Open Access</a>	Select an action ▾
 <a href="#">L2SWBM_Model.zip</a>	2020-06-05	2020-06-05	43.8 KB	<a href="#">Open Access</a>	Select an action ▾
 <a href="#">L2SWBM_input.zip</a>	2020-06-05	2020-06-05	574 KB	<a href="#">Open Access</a>	Select an action ▾
 <a href="#">output_plot_prior.zip</a>	2020-06-05	2020-06-05	2.04 MB	<a href="#">Open Access</a>	Select an action ▾
 <a href="#">output_plot_preview.zip</a>	2020-06-05	2020-06-05	691 KB	<a href="#">Open Access</a>	Select an action ▾
 <a href="#">output_plot_posterior.zip</a>	2020-06-05	2020-06-05	834 KB	<a href="#">Open Access</a>	Select an action ▾


Download All Files (To download individual files, select them in the "Files" panel above)

[Download Zipped Dataset \(4.41 MB in 7 files\)](#)

Best for data sets < 3 GB. Downloads all files plus metadata into a zip file.



# (R) Request

- Data appraisal conversations with researchers
  - Questions about the dataset
  - Pull from the Check and Understand steps
  - Arrive at a shared understanding of the curation process with the researcher (good for future conversations!)
- Record all questions/concerns (ethics, missing files, excessive output files, bad naming conventions, etc.) in your Curator Log 

---

```
Curation log for: <title of dataset>
Handle:
Corresponding researcher name and email:
Curator:
Metadata log created:

*****
Files received:
*****

*****
Changes made to files:
*****

*****
Metadata Changes
*****


*****
Correspondence Notes
*****

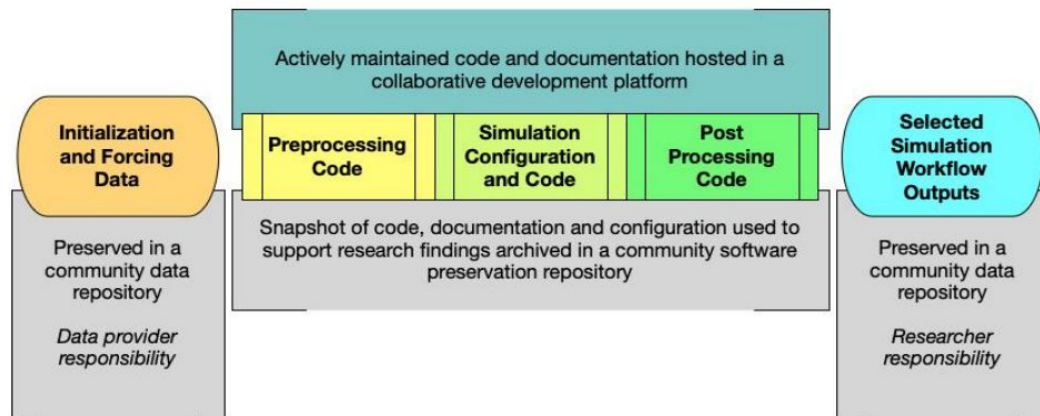
*****
Other issues
*****

*****
Original Metadata from Author:
*****
```

# (R) Request: EarthCube rubric

## EarthCube rubric tool to:

- Understand researcher perspective
- Identify questions for researchers 
- Determine the Use Case for output data to preserve
  - Use Case 1: “Preserve Few Simulation Workflow Outputs”
  - Use Case 2: “Preserve Selected Simulation Workflow Outputs”
  - Use Case 3: “Preserve the majority of simulation workflow outputs”




# (R) Request: Role playing activity

- In groups of 4-6 people:
  - Open the README file from Example Dataset B and review it. Pay particular attention to the “Research Overview” and the list of files under “Files Contained Here.”
  - In your group, decide who will act as the **researcher** depositing the output data, and who will act as the **data curator** receiving the data files.
- Researchers work together and data curators work together. Open the EarthCube rubric and answer the questions as best you can with the Example Dataset B in mind. Answer the questions as you would in your role as the researcher or data curator.
  - You can use the file list in the README file, and/or download the files to explore them further.
  - Note which questions are easier or more difficult for you to answer in the role that you are in.
  - Discuss what questions you would ask the other role (researcher or data curator).
- After completing the rubric, researchers and data curators discuss together:
  - Did you arrive at the same Use Case?
  - What questions were easy or difficult to answer?
  - What questions do you have for each other?

# (A) Augmenting for discoverability

**Review information** received from the researcher from initial deposit and all subsequent conversations to update metadata and documentation as appropriate:

- Facilitate **discoverability**:
  - Add links to related publications, grants, reports, source data, etc.
  - Provide additional description of files as appropriate for external indexing or other purposes.
  - Add subject terms
- Ensure **keywords** are sufficient and representative
- Record all **changes** in the Curator Log 
- Provide suggestions to improve **accessibility** of content (e.g., alt-text or additional descriptions; color contrast; etc)

# (A) Augment resources

## Metadata Standards Catalog

The RDA Metadata Standards Catalog is a collaborative, open directory of metadata standards applicable to research data. It is offered to the international academic community to help address infrastructure challenges.

[Read more details about the scope of the Catalog](#)[Read our terms of use](#)[Read our accessibility statement](#)[Contribute to the Catalog](#)[Explore our API](#)

## Metadata standards, profiles and schemes

[Browse by scheme name](#)[Browse by subject](#)[Search](#)

## “Standards”

- [Metadata Standards Catalog \(RDA\)](#)
  - [Index of subjects](#)
- [Licenses](#)

## Metadata-related entities

[Browse mappings between schemes](#)[Browse metadata-related tools](#)[Browse organizations](#)[Browse funders](#)[Browse organizations that maintain schemes](#)[Browse known users of schemes](#)[Browse organizations that have endorsed schemes](#)[Browse endorsements](#)

# (T) Transform

In this step, consider the file formats in the dataset to make them more interoperable, reusable, preservation friendly, and non-proprietary when possible. See Transform slide (slide 46) for a list of recommended file formats.

- Check, Understand, and Augment steps lay the foundation for Transform step.
- Opportunity for Transform step exists in all the Common elements of simulation models (slide 13): Input, code/software, output, and documentation. However, the most common and less risky transformation is in the output data and documentation. There is a high risk of making mistakes, breaking the workflow, or data corruption in attempting to transform the input and code. For code/software suggestions can be made for open source alternatives.

# (T) Transform: Key ethical considerations


- Decide how to balance the potential benefits of transformation with the risks of mistakes and loss of content/context, especially if the curator or repository will be performing transformation. Document the decision.
- Except in situations where there is extremely low or no risk to cause any unintended data corruption, it is recommended that transformation be performed by the researcher.
- It is important to reach a shared understanding with the researcher of who is responsible for transformation tasks (curator or researcher).

# (T) Transform: Common steps

- Identify specialized file formats and their restrictions
  - Example: Is the software freely available? If so, link to it or archive it alongside the data
  - Example: convert .docx README file to plain text file or pdf file (.pdf) format
- Propose open source or more reusable formats when appropriate
  - Example: convert Excel to CSV
- Retain original file formats



# (T) Transform: Essential tasks

- Check whether preferred file formats are in use
  - If not, recommend conversion
  - Retain original formats
- Check whether software needed is readily available
  - Suggest open source alternatives, if applicable and appropriate (Example: Proprietary Tecplot vs Open-source Paraview or VisIt)
  - Ensure software and software version is documented
- Update curator log 

# Activity 3: Transformation

**Activity:** Demonstrate data inspection, tool search, and transformation process to make them more interoperable, reusable, preservation friendly, and non-proprietary.

- **Example Dataset C**

Horna Munoz, Daniel; Constantinescu, George; Rhoads, Bruce ; Lewis, Quinn; Sukhodolov, Alexander (2020):

Confluence Density Effects Simulation Database. University of Illinois at Urbana-Champaign.

[https://doi.org/10.13012/B2IDB-6257171\\_V1](https://doi.org/10.13012/B2IDB-6257171_V1)

## Confluence Density Effects Simulation Database

### Citation:

Horna Munoz, Daniel; Constantinescu, George; Rhoads, Bruce ; Lewis, Quinn; Sukhodolov, Alexander (2020): Confluence Density Effects Simulation Database. University of Illinois at Urbana-Champaign. [https://doi.org/10.13012/B2IDB-6257171\\_V1](https://doi.org/10.13012/B2IDB-6257171_V1) [Export Citation](#)

### Persistent link for this dataset:

[https://doi.org/10.13012/B2IDB-6257171\\_V1](https://doi.org/10.13012/B2IDB-6257171_V1)

### Dataset Description

This data set shows how density effects have an important influence on mixing at a small river confluence. The data consist of results of simulations using a detached eddy simulation model.

### Subject

[Physical Sciences](#)

### Keywords

confluence; flow dynamics; density effects

### License

CC0

### Funder

U.S. National Science Foundation (NSF) - **Grant:** BCS 1359911

### Corresponding Creator

Daniel Horna Munoz

# (T) Transform: Files for transformation

1. **Output files:** transform the two Excel (.xls) files: **Figure17ab-mixing metric - Quinn - all.xls** and **Figure18abvolume of mixed fluid Cm - 0point2.xls** to Comma Separated Value (.csv) files. Also consider removing the spaces in the file names
2. **Documentation files:** Transform the Microsoft Word, **EXPLANATION OF FILES.docx** to plain text (.txt) or pdf format
3. Consider removing the spaces in the file names

# (T) Transform: Open file formats

- What is the .lay file? Can it be transformed into an open file format? suggest alternative open format and conversion tools to the proprietary **Tecplot**, <https://www.tecplot.com/products/tecplot-360/>
- Alternatives to Tecplot (computational fluid dynamics 3d simulation & visualization software)
  - **Paraview** (free and open-source), recommended for Windows platform: <https://www.paraview.org/>
  - **OpenFOAM** (free and open-source), mainly for unix based platforms: <https://www.openfoam.com/>
  - **Visit** (free and open-source): <https://visit-dav.github.io/visit-website/about/>
- Evaluate: "Check that any transformations didn't introduce problems"

# (T) Transform: Data formats

For a list of recommended Digital Data Formats, you may refer to the following sources, among others:

- University of Virginia:  
<https://data.library.virginia.edu/data-management/plan/format-types/>
- Library of Congress:  
<https://www.loc.gov/preservation/resources/rfs/TOC.html>
- Cornell University: <http://guides.library.cornell.edu/ecommons/formats>

# (T) Transform: Summary

- **Convert** any data visualization(s) that are not accessible (e.g., R visualizations, which need to be converted for screen reader use, or visualizations that do not meet color contrast guidelines)
- **Reorganize** files as appropriate
- **Standardize** file names (eg., remove spaces from file names)
- **Record** any transformations in Curator Log: Transformation steps and who made the transformation should be well documented. Be aware of what is transformable and what's not.
- **Evaluate**: "Check that any transformations didn't introduce problems"

# Workshop Summary

In this workshop we used the CURATE(D) framework to engage with real-world datasets:

- Gaining a basic understanding of simulation-based research
- Learning how to recognize data and code associated with simulation research
- Knowing what kinds of curation questions to ask of researchers about their simulation-based datasets
- Understanding related curation resources

These slides and additional instructional materials will be made available on the DCN website: <https://hdl.handle.net/11299/202807>

# Wrap-up activity

What did you learn and what is still muddy?

<https://www.menti.com/XXXX>

Menti.com code XX XX X





# Thanks for your participation!!!

Thanks for joining us on this  
simulation research curation  
roller coaster!

