EarthCube RCN -Determining Best Practices for Preservation and Replicability of Model Data

Doug Schuster, NCAR
Matt Mayernik, NCAR
Gretchen Mullendore, NCAR/U. North Dakota

https://modeldatarcn.github.io/

NSF Awards #1929773, #1929757

Project Motivation –Open Data Access Expectations

- Evolving community open access expectations have led to data management requirements from funding agencies and publishers
  - Data management requirements for simulation output, however, have not been clear

https://modeldatarcn.github.io/

Project Steering Committee

- **Adam Clark**, NOAA/University of Oklahoma
- **Laura Condon**, University of Arizona, Hydrology and Atmospheric Sciences
- **Gokhan Danabasoglu**, NCAR, Climate and Global Dynamics Laboratory
- **Josh Hacker**, Jupiter
- **Michael A. Friedman**, American Meteorological Society (AMS)
- **Cathy Smith**, NOAA, Physical Sciences Laboratory
- **Gary Strand**, NCAR, Climate and Global Dynamics Laboratory

Student members:

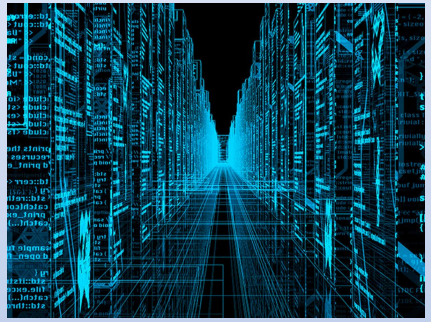
- **Jared Marquis**, University of North Dakota, Atmospheric Sciences
- **Elisa Murillo**, University of Oklahoma, School of Meteorology

https://modeldatarcn.github.io/

Project Motivation -What to do about model data?

We know the answer is not “preserve all the data/output for all projects”

- Too expensive due to large data volumes
- Not all model outputs are relevant to the research topic



https://modeldatarcn.github.io/

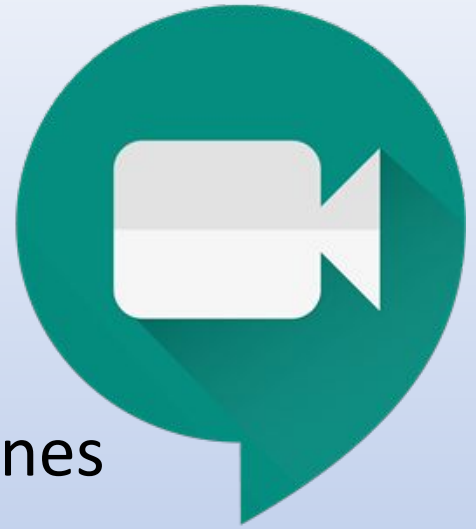
Project Motivation -What to do about model data?

RCN Project -Bring together a diverse group of modeling experts to develop simulation output preservation guidance:

1. Develop a rubric to guide researchers in determining what model output to preserve and share to communicate knowledge
2. Refine rubric with extensive set of use cases
3. Disseminate best practices document to broader community

https://modeldatarcn.github.io/

Project Activities -Workshops



Workshop #1 - May 5-8, 2020 - 45 participants
Workshop #2 - Aug. 3-6, 2020 - 40 participants

- Participants:
  - Experienced modelers from a wide range of disciplines
  - Data and technology experts
  - Publishers, editors
  - Inclusion of advanced graduate students and early career scientists
- Develop draft rubric
- Develop draft use cases according to rubric score
- Discuss challenges in achieving simulation workflow components

https://modeldatarcn.github.io/

Project Engagement

Conferences:

- AMS (Town Halls), AGU, EarthCube
- Earth Science Information Partners (ESIP) Meeting
- Community Earth System Model (CESM) Annual Meeting
- NASA Physical Oceanography DAAC Users Meeting
- US CLIVAR Inter-Agency Group Meeting
- Coalition for Publishing Data in the Earth and Space Sciences (COPDESS)

Collaborations:

- AGU Publishing, AMS Publishing
- Use of rubric by institutions (e.g., NCAR) and industry (e.g., AER)

https://modeldatarcn.github.io/

Rubric Overview

https://modeldatarcn.github.io/

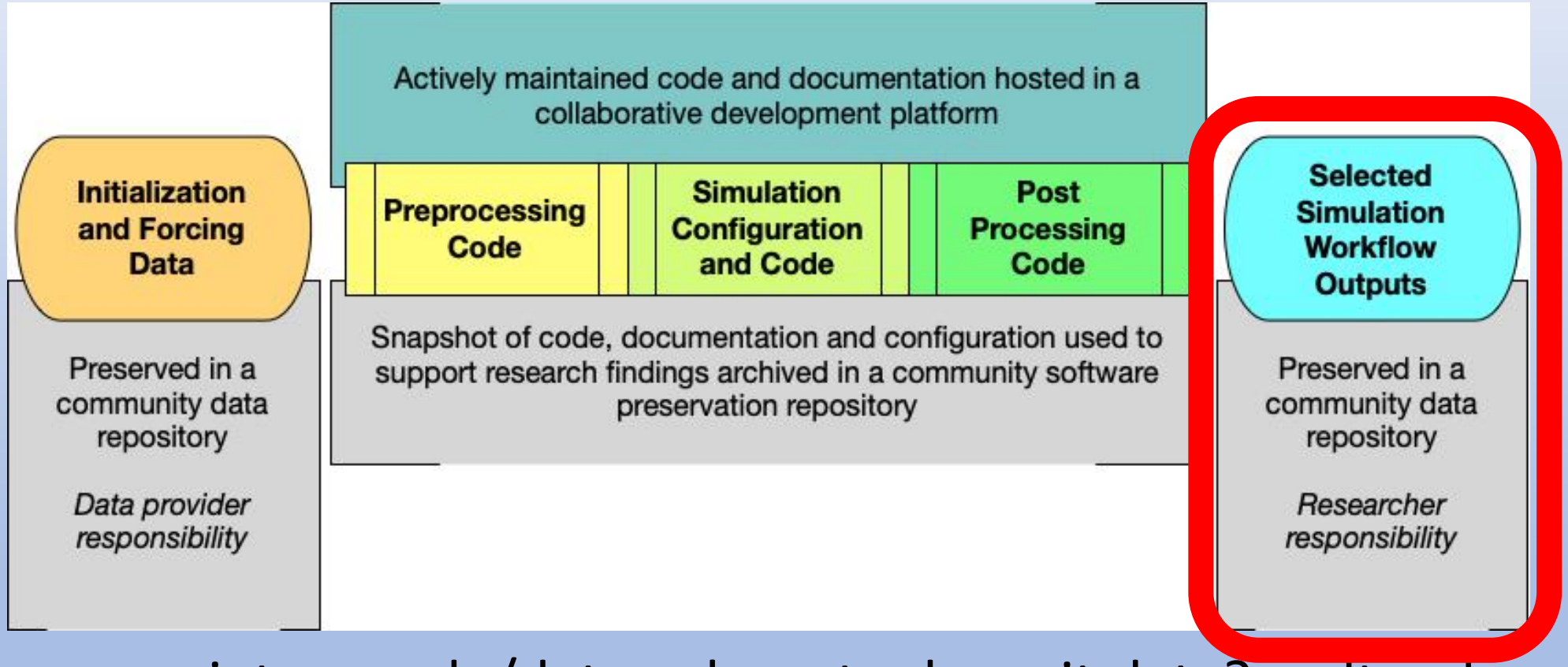
Rubric -“What Model Output to Preserve?”

To assist a researcher in determining what simulation outputs should be deposited in a trusted community repository to communicate knowledge.

https://modeldatarcn.github.io/

-What to preserve and share for all projects?

Open science expectations for simulation based research. *Frontiers in Climate*, 2021. https://doi.org/10.3389/fclim.2021.763420



barriers: proprietary code/data, where to deposit data?, cultural resistance

https://modeldatarcn.github.io/

Rubric Structure

Simulation Descriptor Theme					

https://modeldatarcn.github.io/

Rubric Structure

Simulation Descriptor Theme					
Big Picture Question					

https://modeldatarcn.github.io/

Rubric Structure

Simulation Descriptor Theme					
Big Picture Question	Simulation Descriptors				
	Descriptor	Descriptor definition			

https://modeldatarcn.github.io/

Rubric Structure

Simulation Descriptor Theme					
Big Picture Question	Simulation Descriptors		Simulation Descriptor Classes		
	Descriptor	Descriptor definition	Class 1 Preserve less output	Class 2 Preserve some output	Class 3 Preserve more output

https://modeldatarcn.github.io/

Cost

Is it more costly to rerun a full simulation workflow or preserve model output products in a community repository?

https://modeldatarcn.github.io/

Cost

Is it more costly to rerun a full simulation workflow or preserve model output products in a community repository?

Simulation Descriptor Themes:

- **Cost of Running Simulation Workflow**
  - *What is the cost to produce your simulation workflow outputs?*

https://modeldatarcn.github.io/

Cost

Is it more costly to rerun a full simulation workflow or preserve model output products in a community repository?

Simulation Descriptor Themes:

- **Cost of Running Simulation Workflow**
  - *What is the cost to produce your simulation workflow outputs?*
- **Repository Data Management Services Cost**
  - *What is the cost for you to archive the output in a community repository to preserve and provide access to your simulation workflow outputs for a minimum period of time?*

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow

Big Picture Question					
What is the cost to produce your simulation workflow outputs?					

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow

Big Picture Question	Simulation Descriptors		Simulation Descriptor Classes		
	Descriptor	Descriptor definition	Class 1 Preserve less output	Class 2 Preserve some output	Class 3 Preserve more output
What is the cost to produce your simulation workflow outputs?					

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow

Big Picture Question	Simulation Descriptors		Simulation Descriptor Classes		
	Descriptor	**Descriptor definition**	**Class 1** Preserve less output	**Class 2** Preserve some output	**Class 3** Preserve more output
What is the cost to produce your simulation workflow outputs?	**Computational Cost of Running the Simulation Workflow**	The economic cost (combination of run time and computer access costs) of completing simulation workflow	Small computational cost and no special platform needs	Moderate computational cost, but access to needed platforms straightforward	High computational cost. Need a large and /or specialized compute capability...

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow

Big Picture Question	Simulation Descriptors		Simulation Descriptor Classes		
	Descriptor	**Descriptor definition**	**Class 1** Preserve less output	**Class 2** Preserve some output	**Class 3** Preserve more output
What is the cost to produce your simulation workflow outputs?	**Computational Cost of Running the Simulation Workflow**	The economic cost (combination of run time and computer access costs) of completing simulation workflow	Small computational cost and no special platform needs	Moderate computational cost, but access to needed platforms straightforward	High computational cost. Need a large and /or specialized compute capability...
	Human Resource cost of producing the simulation workflow	Person hours required to reproduce a simulation dataset	Trivial effort required to replicate simulation for most end users		Significant time & expertise required to replicate simulation...

https://modeldatarcn.github.io/

Section Theme: Repository Data Management Services Cost

Big Picture Question	Simulation Descriptors		Simulation Descriptor Classes		
	Descriptor	Descriptor definition	Class 1 Preserve less output	Class 2 Preserve some output	Class 3 Preserve more output
What is the cost for you to archive the output in a FAIR aligned community repository..?					

https://modeldatarcn.github.io/

Section Theme: Repository Data Management Services Cost

Big Picture Question	Simulation Descriptors		Simulation Descriptor Classes		
	Descriptor	**Descriptor definition**	**Class 1** Preserve less output	**Class 2** Preserve some output	**Class 3** Preserve more output
What is the cost for you to archive the output in a FAIR community repository..?	**Repository Supported Data Curation Cost**	The economic cost of curating simulation output in a community repository, for a minimum time period.	Community repository data curation expenses are prohibitive due to large volume of the expected model outputs.		Would be inexpensive to curate the complete simulation workflow output for a minimum number of years in a community repository.

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

- **Community Commitment** - Is it anticipated that your simulation workflow outputs will have broad community impact and downstream reuse?
- **Research Workflow Accessibility** - Would it be reasonable to expect others in your academic discipline to rerun your full simulation workflow?
- **Data Accessibility** - Would it be reasonable to expect others to access and use simulation workflow outputs?
- **Research Feature Reproducibility** - Are physical features generated by a simulation reproducible?
- **Cost** - Is it more costly to re-run a full simulation workflow or preserve model output products in a FAIR aligned repository?

https://modeldatarcn.github.io/

Rubric -Total Score of Descriptor Section Themes

Rubric Total Raw Score. (Min=17, Max=51)	1	Rubric Total Weighted Score. (Min=17, Max=90)	1

	Rubric Total Weighted Score < 48	48 <= Rubric Total Weighted Score <= 72	72 < Rubric Total Weighted Score
	Preserve few simulation workflow outputs	Preserve selected simulation workflow outputs	Preserve the majority of simulation workflow outputs
	Preserve and provide access to simulation workflow configuration and code components	Preserve and provide access to simulation workflow configuration and code components	Preserve and provide access to simulation workflow configuration and code components
	See Use Case 1	See Use Case 2	See Use Case 3

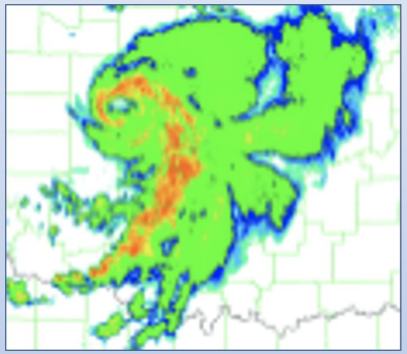
https://modeldatarcn.github.io/

Reference Use Cases and Emerging Ideas

Matt Mayernik

Reference Use Case Examples -Use Case 1

Preserve Few Simulation Workflow Outputs (Score < 48)



- *Semi-idealized WRF-ARW-based numerical simulations of tropical cyclones over land.*
  - Idealized Process Study – Goal is knowledge production
  - Preserve and share: input data and simulation configuration and codes

https://modeldatarcn.github.io/

Reference Use Case Examples -Use Case 2

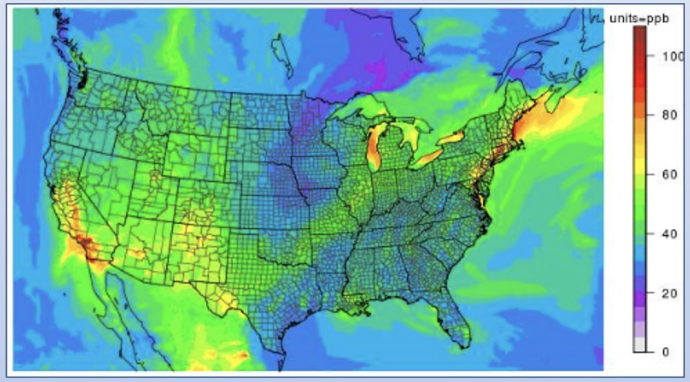
Preserve Selected Simulation Workflow Outputs (48 <= Score <= 72)

- *Warn-on-Forecast - an on-demand convection-allowing ensemble forecast system*
  - Preserve and share: input data, simulation configuration and codes, a portion of the processed model output
  - Important environmental fields are saved in the form of “summary files”, which are a fraction of the raw output

https://modeldatarcn.github.io/

Reference Use Case Examples -Use Case 3

Preserve the Majority of Simulation Workflow Outputs (Score > 72)



- *Modeling ammonia in the atmosphere*
  - Use input obs from NASA/NOAA and a series of model steps to produce ammonia emission profiles. Goal is data production
  - Preserve and share: simulation configuration and codes, and processed model output related to ammonia

https://modeldatarcn.github.io/

Use Case Compilation

- How do scientists make decisions about what data/files to deposit in a data repository?
- Went through use case template with 9 members of NCAR staff who have deposited data in a repository previously
- High level takeaways
  - Main purpose for depositing data - To fulfill publisher requirements
  - Other purposes - to provide data for some community, reduce work for others (for complex models)
  - Data collections varied considerably - Some were long lists of a single file type of the same size, others included multiple file types of varying sizes
  - Rubric score generally was consistent with their choices about what to deposit in the repository

Collaborator: David Eby (University of Denver / University of Illinois at Urbana-Champaign)

Use Case Outcomes - Varying Practices

1. Input data - little consistency in whether people include input data in their data collection or not

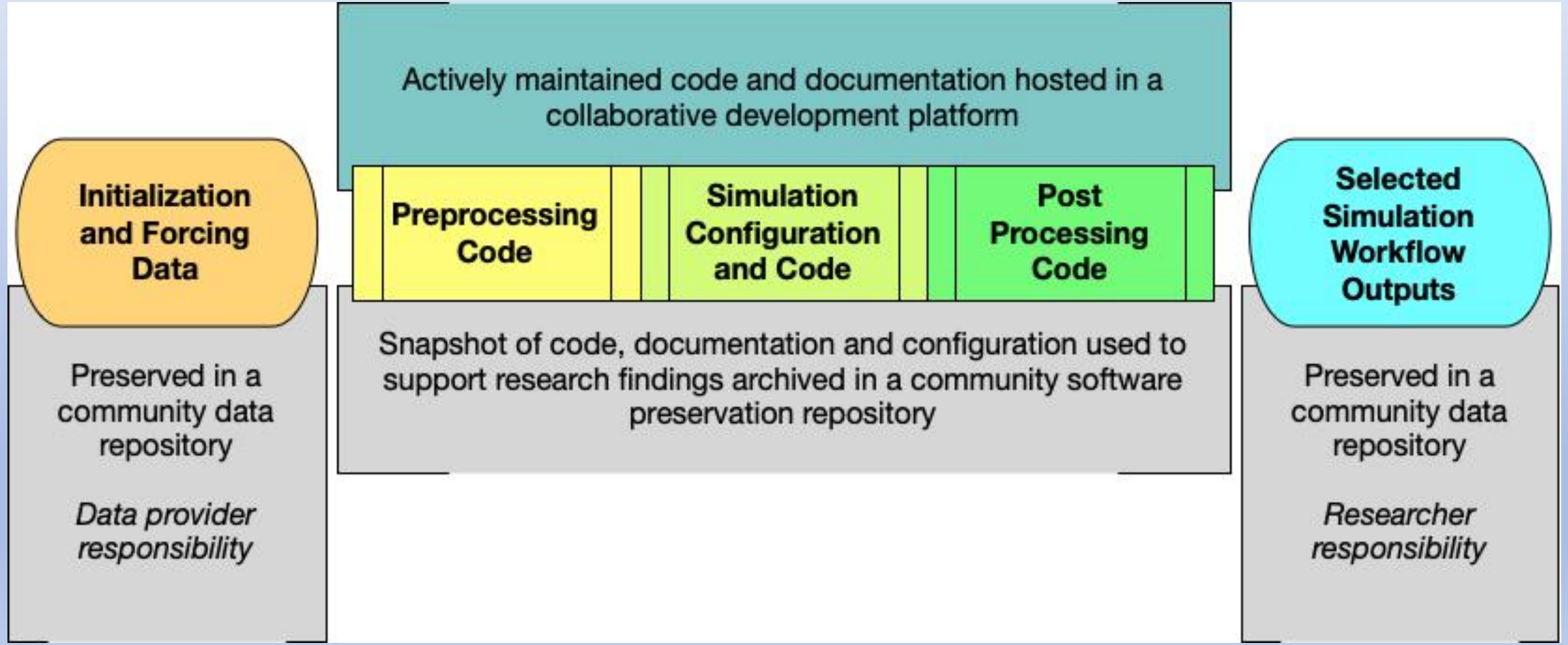
2. Software - varying practices about whether software are shared. It was more common for model code to be openly available than pre/post processing code

3. Model outputs - it was more common for people to have deposited some subset of their model output, vs. depositing everything.

4. Documentation - documentation is commonly missing, or highly variable in extent.

Collaborator: David Eby (University of Denver / University of Illinois at Urbana-Champaign)

Open science expectations for simulation based research. *Frontiers in Climate*, 2021. https://doi.org/10.3389/fclim.2021.763420



Replicability vs. Reproducibility

- The primary goal in earth science is replicability, not computational reproducibility. [1]
- Provide enough information about the workflow and selected derived outputs to communicate the important environmental characteristics to allow a future researcher to build off of the original study.
- For highly nonlinear simulation studies, computational reproducibility should not be expected, nor is it needed. Findings that only work when bit-reproducibility is needed are problematic for others to build on. [2]

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

[2] Bush, Rosemary, A. Dutton, M. Evans, R. Loft, and G.A. Schmidt. (2020). "Perspectives on Data Reproducibility and Replicability in Paleoclimate and Climate Science." *Harvard Data Science Review,* 2(4). https://doi.org/10.1162/99608f92.00cd8f85

How should preservation of model software and outputs differ for projects that are oriented toward knowledge production vs projects oriented toward data production?

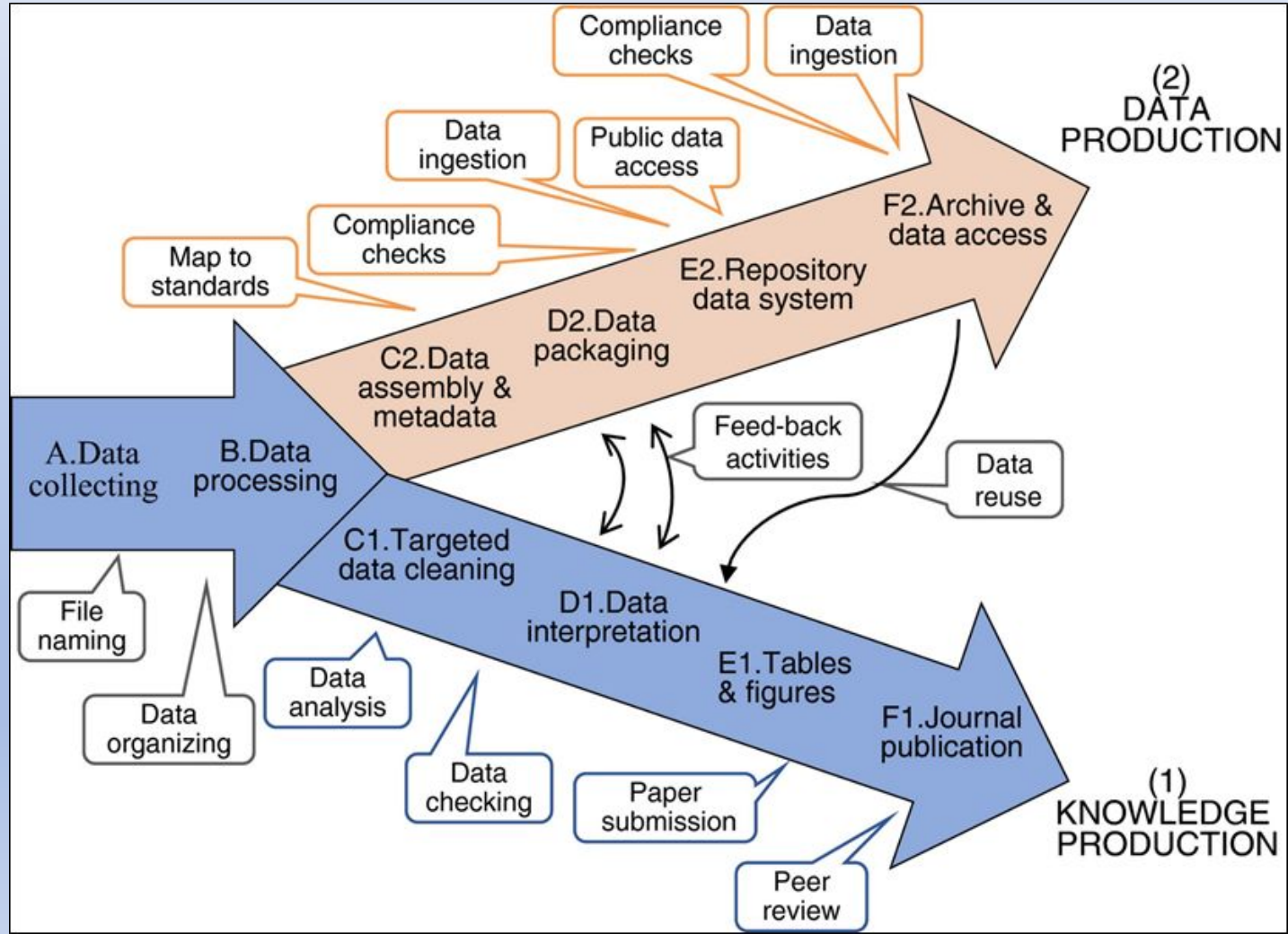


Figure from:
Baker, K.S. & Mayernik, M.S. (2020). Disentangling knowledge production and data production. *Ecosphere*, 11(7). https://doi.org/10.1002/ecs2.3191

Interdependency with Technologies

- Improved technological capabilities, including cloud storage, are critical to dealing with model data, but they do not solve all data preservation needs.
- Without data stewardship and curation, cloud storage is nothing more than a modern version of "anonymous FTP".
- Packages like Jupyter are good for transparency and reproducibility but not good for curation.
- Without investment in data curation personnel, the potential benefits of improved technological capabilities will not be realized.

https://modeldatarcn.github.io/

https://thenounproject.com/term/money-cloud/66766/

What curation support is needed to enable sharing and preservation for geoscience simulation models and their output?

- Researchers are currently spending a significant portion of their own time dealing with data curation.
- The ecosystem of community repositories to support Atmospheric Science is sparse.
- We need a coordinated effort to fund personnel to assist researchers in data curation, as well as investment in the needed repository preservation and stewardship services.
- Potential ways forward [3]:
  - 1) augment existing geoscience data repositories to scale up their capacity
  - 2) identify non-specialized data repositories that fulfill open access objectives
  - 3) develop a data repository liaison service
  - 4) create new data repository services

[3] Mayernik, M.S, D. Schuster, S. Hou, & G.J. Stossmeister. (2018). *Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop Report*, NCAR/TN-552+PROC. Boulder, CO: National Center for Atmospheric Research. https://doi.org/10.5065/D6NC601B

What cultural barriers impede geoscience modelers from making progress on these topics?

- Researchers need to be rewarded for collaboration, not data/software hoarding.
- Reward good data and software sharing practices in addition to good publications.
- Withholding data and software perpetuates inequalities and limits scientific opportunities.
- Equity issues in preventing access to data and software for other people who can't compile the data themselves (not enough storage or network bandwidth) or who don't have existing relationships with the authors of an article.

https://modeldatarcn.github.io/

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

Emerging Ideas Summary

- Research that is primarily oriented toward **“knowledge production”** should preserve minimal model output in repositories. **“Data production”** oriented research should include appropriate resources to support anticipated data preservation and community data access needs.

- A **coordinated effort is needed to support personnel to assist researchers in data and software curation**, as well as investment in the needed repository preservation and stewardship services

- **Cultural barriers** impede modelers from embracing **open software and open data**

https://modeldatarcn.github.io/

NSF EarthCube Transforming Geosciences Research

Findings from Workshop 3

- **Sustainable Curation**
  - Software and data management plans need to be well thought out by PIs/creators and elevated in importance by funding agencies (broader impact).
  - Funding should come from agencies specifically for data/software management needs
  - Incorporate training for data and software management in standard curriculum
- **Determining Lifetime for Simulation Data**
  - Simulation data do not need to be preserved indefinitely
  - Plan and advertise de-accession strategy at the point when data is deposited
  - Use a defined process to evaluate when simulation data can be purged from a repo

https://modeldatarcn.github.io/

NSF EarthCube Transforming Geosciences Research

Findings from Workshop 3

- **Incentivizing Data and Software Preservation and Sharing**
  - Showcase open science based research success stories
  - Update promotion and tenure process to support sharing of code and data
  - Raise the visibility of open science achievements -publisher and societal awards
- **Equitable Access to Data and Software Curation and Analysis Resources**
  - Provide the resources for under resourced communities to meet open science expectations
    - Access to data proximate compute and trusted data/software repositories
    - Accessible training and support: “National virtual data curation laboratory”
  - Invest in building relationships

https://modeldatarcn.github.io/

NSF EarthCube Transforming Geosciences Research

Next Steps https://modeldatarcn.github.io

- Continue to engage publishers, sponsors, and professional societies
  - Town hall meeting at AMS 2023 annual meeting
- Summarize and add additional use case examples
- Publish project outcomes

https://modeldatarcn.github.io/

NSF EarthCube Transforming Geosciences Research