# 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

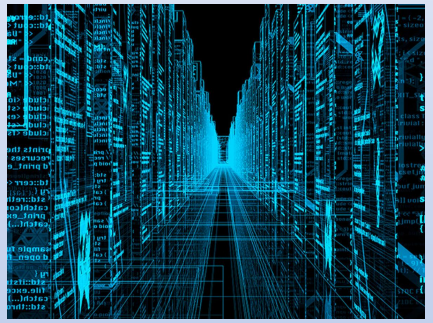
- Evolving community open access expectations have led to data management requirements from funding agencies and publishers
  - Data management requirements for simulation output, however, have not been clear

https://modeldatarcn.github.io/

# Project Motivation -What to do about model data?

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

- 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 comprehensive list of simulation descriptors
2. Use simulation descriptors to build rubric for output preservation classes
3. Refine rubric with extensive set of use cases
4. Disseminate best practices document to broader community

https://modeldatarcn.github.io/

# Project Overview

- 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 Activities -Workshops

- **Workshop #1** - May 5-8, 2020 - 45 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
  - Brainstorm and prioritize simulation descriptors
  - descriptor list: 102 -> 17
  - organized by theme
  - Product: First draft of model rubric

https://modeldatarcn.github.io/

# Preserve for All Projects

Unless already publicly available from another provider:

- simulation code
- initialization data
- simulation setup (e.g., parameterization selection)
- pre-processing code
- post-processing code

barriers: proprietary code/data, access issues, cultural resistance

https://modeldatarcn.github.io/

# Rubric -“What Model Output to Preserve?”

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

https://modeldatarcn.github.io/

# Rubric -Simulation/Experiment Descriptors and Descriptor Classes

PRESERVE LESS OUTPUT ⟷ PRESERVE MORE OUTPUT

| Descriptor | Descriptor definition | Class 1 | Class 2 | Class 3 |
|---|---|---|---|---|
| **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 compute capability and/or can only be produced with specialized platforms |
| **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 |

https://modeldatarcn.github.io/

# Project Activities -Workshops

- **Workshop #2** - Aug. 3-6, 2020 - ~40 participants
  - Refine rubric with use case testing
  - Participants brought cases from own experience to test rubric
  - Discussed emerging ideas
  - Product #1: Examples for different use case types
    - Low rubric score: Preserve few outputs
    - Middle: Preserve selected outputs
    - High: Preserve majority outputs
  - Product #2: Use case template
    - Building out use case set
    - Working on new "Broader Impacts" section to encourage data sharing decisions be more inclusive

https://modeldatarcn.github.io/

# Emerging Ideas

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

**Replicability, not reproducibility,** is the goal of most research

**“Knowledge production”** research should preserve minimal model output in repositories; **“Data production”** research should include appropriate resources to support anticipated data preservation and community data access needs

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

**No individual technology**, e.g. cloud storage, will solve all our data needs

https://modeldatarcn.github.io/

NSF EarthCube Transforming Geosciences Research

# Rubric and Reference Use Cases

Doug Schuster

https://modeldatarcn.github.io/


# Rubric -“What Model Output to Preserve?”

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

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<br>Preserve less output | Class 2<br>Preserve some output | Class 3<br>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 FAIR aligned repository?

https://modeldatarcn.github.io/

# Cost

Is it more costly to rerun a full simulation workflow or preserve model output products in a FAIR aligned 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 FAIR aligned 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 FAIR aligned 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<br>Preserve less output | Class 2<br>Preserve some output | Class 3<br>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<br>Preserve less output | Class 2<br>Preserve some output | Class 3<br>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 aligned 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** - Does a project fall into the “Data Production” or “Knowledge Production” category?
- **Research Workflow Accessibility** - Would it be reasonable to expect others with discipline specific knowledge to rerun a 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

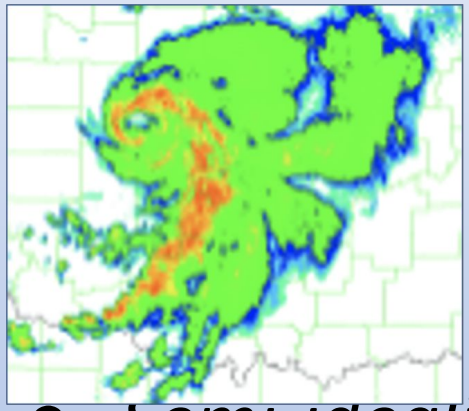
| Rubric Total Raw Score. (Min=17, Max=51) | 1 | Rubric Total Weighted Score. (Min=17, Max=90) | 1 |
|---|---|---|---|

| | 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 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

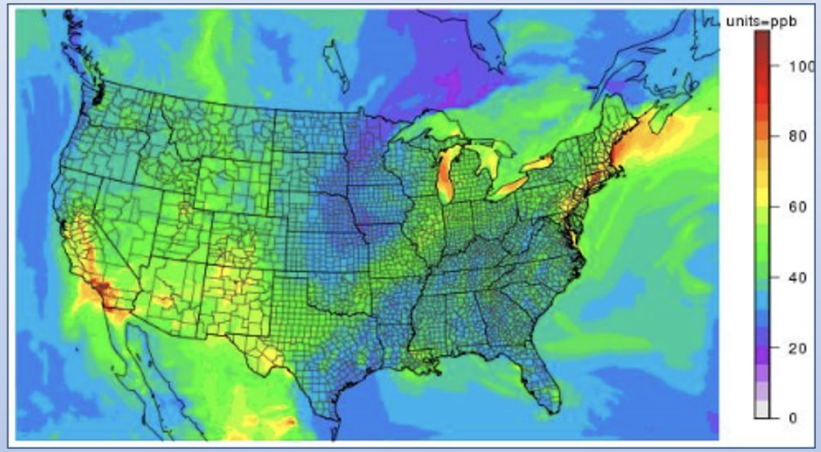
## **Preserve Selected Simulation Workflow Outputs** (48 <= Score <= 72)


- *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/

# Emerging Ideas

Matt Mayernik

# 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?

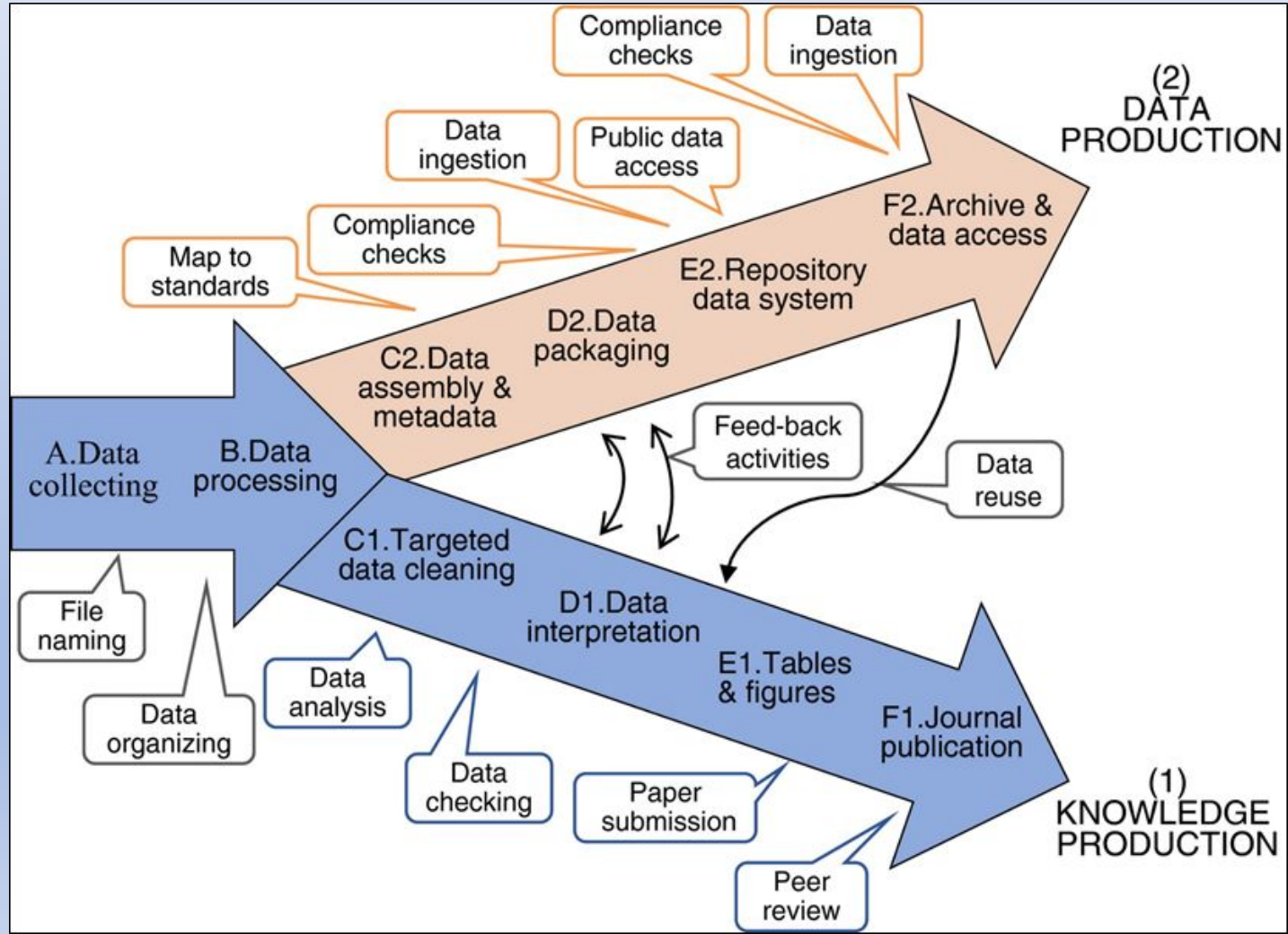



Figure from:
Baker, K.S. & Mayernik, M.S. (2020). Disentangling knowledge production and data production. *Ecosphere*, 11(7). https://doi.org/10.1002/ecs2.3191

# 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

- **Replicability vs. reproducibility** is an important consideration for modelers
- 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.
- **No individual technology**, e.g. cloud storage, will solve all our data 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

# Next Steps https://modeldatarcn.github.io

- Engage publishers, sponsors, and professional societies
- Work with individual volunteers to refine use cases
- 2021 EarthCube Annual meeting
- Additional workshops in 2022
  - Refine rubric and associated documentation further
  - Explore emerging themes
- Publish outcomes

**Contact schuster@ucar.edu if you're interested in joining the modeldatarcn mailing list for updates on RCN activities**

https://modeldatarcn.github.io/

NSF EarthCube Transforming Geosciences Research

# Discussion Questions

- **How could the Rubric best be positioned to inform future development of publisher data management requirements?**
- **How is the best way to proceed without creating unreasonable barriers to publishing?**
- **Other ideas?**

https://modeldatarcn.github.io/

NSF EarthCube Transforming Geosciences Research