Tools to Assist Simulation Based Researchers in Deciding What Project Outputs to Preserve and Share

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

https://modeldatarcn.github.io/

NSF Awards #1929773, #1929757

EarthCube RCN “What about Model data?”, Determining Best Practices for Preservation and Replicability

- **<u>Project motivation:</u>**
- Evolving community open access expectations have led to data management requirements from funding agencies and publishers
  - Data management requirements for simulation output have not been clear

<u>https://modeldatarcn.github.io/</u>

Project Activities -Workshops

Workshop #1 - May 5-8, 2020 - 45 participants
Workshop #2 - Aug. 3-6, 2020 - 40 participants
Workshop #3 - Jul. 25-27, 2022 - 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 rubric
- Develop use cases according to rubric score
- Discuss challenges in achieving data and software management goals

https://modeldatarcn.github.io/

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/

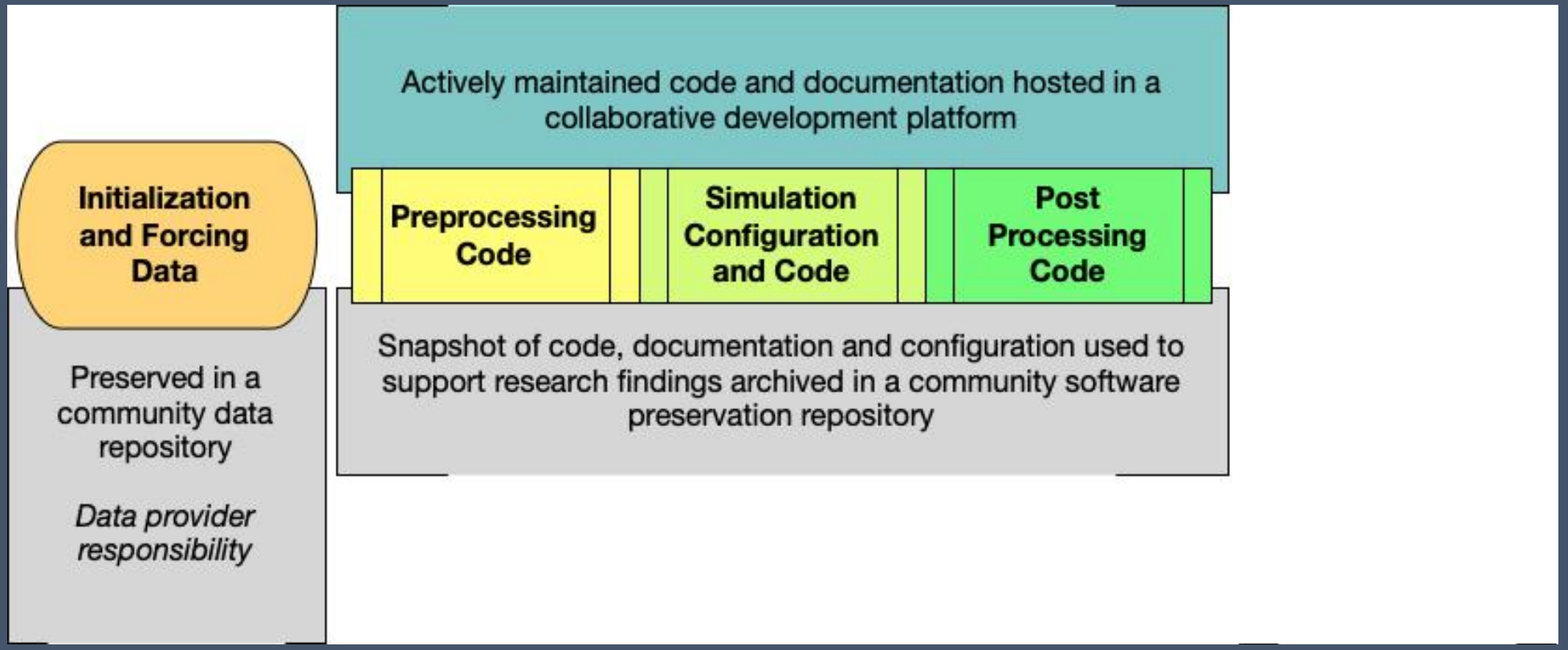
Project findings -What to preserve and share for all projects?

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

https://modeldatarcn.github.io/

Project findings -What to preserve and share for all projects?

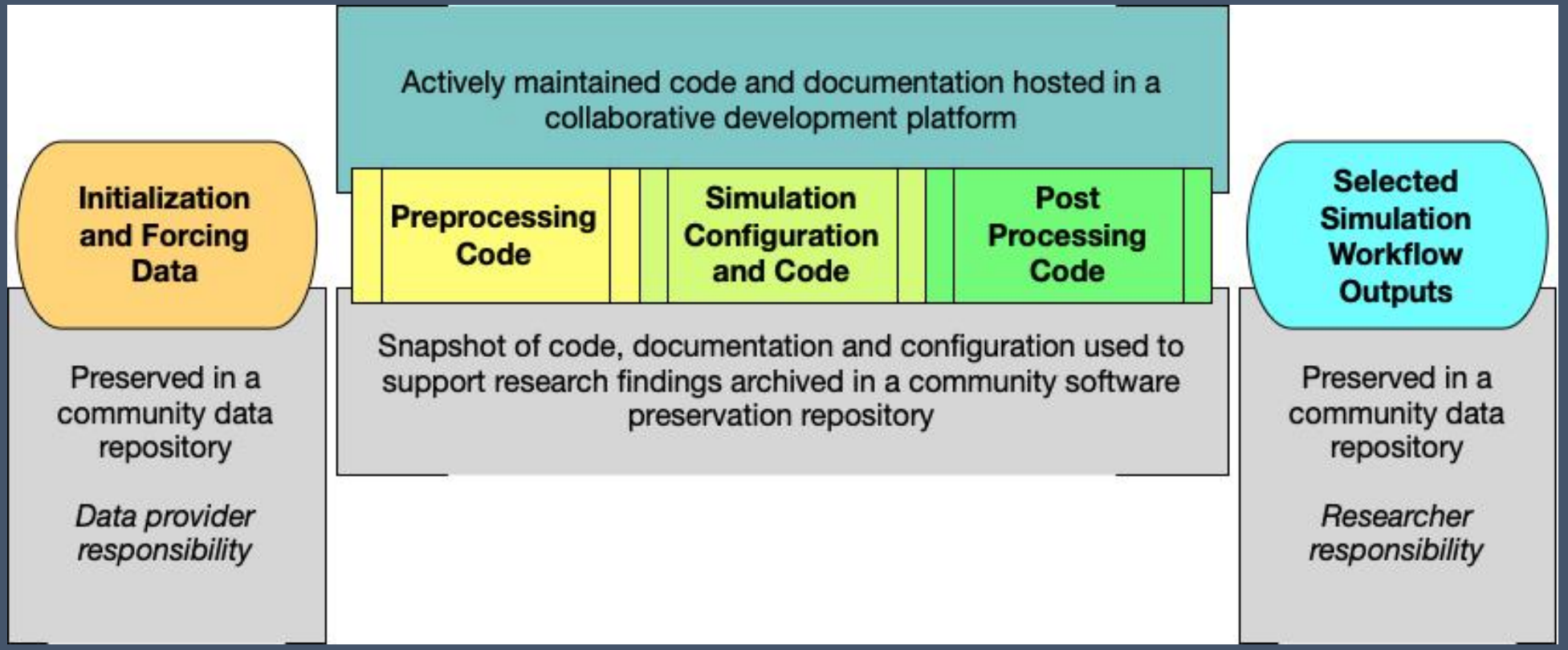
Open science expectations for simulation based research. *Frontiers in Climate*, 2021. https://doi.org/10.3389/fclim.2021.763420



https://modeldatarcn.github.io/

Project findings -What to preserve and share for all projects?

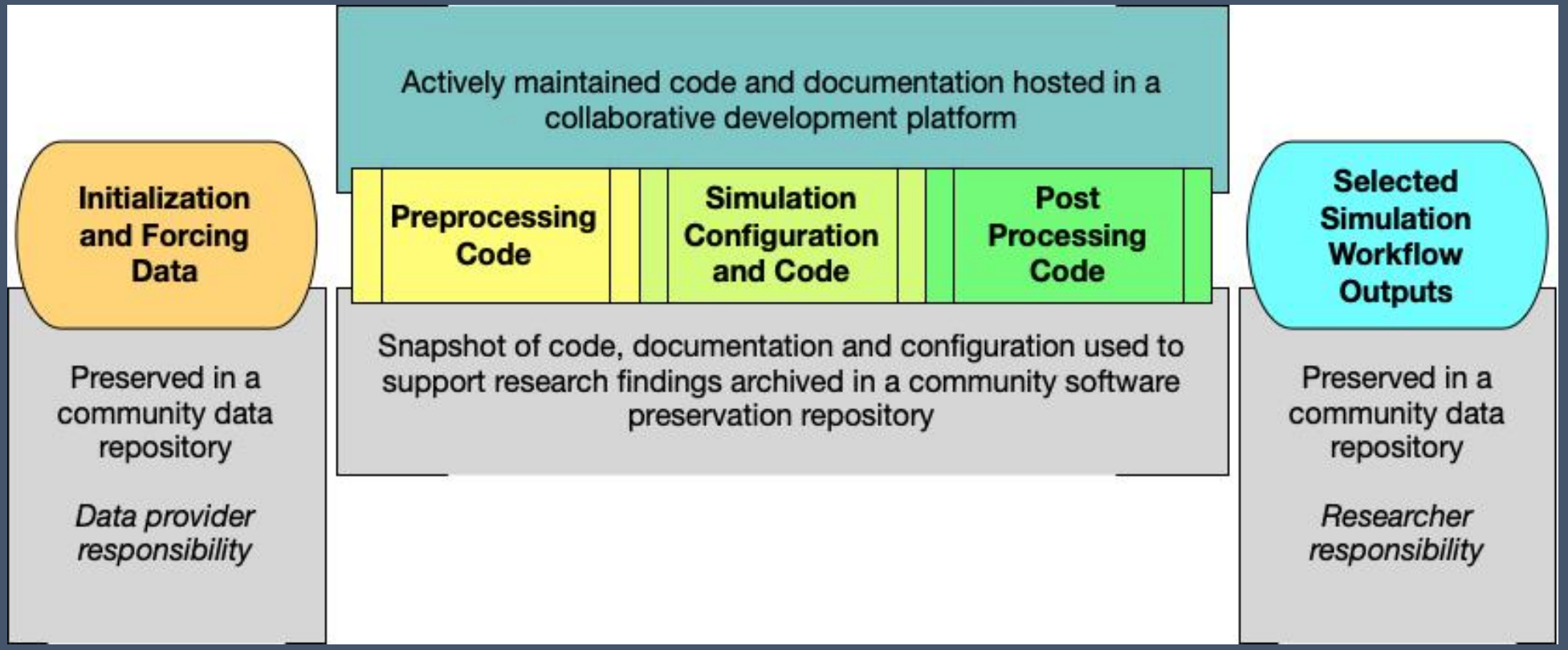
Open science expectations for simulation based research. *Frontiers in Climate*, 2021. https://doi.org/10.3389/fclim.2021.763420



https://modeldatarcn.github.io/

Project findings -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



Use rubric for guidance on what simulation workflow outputs to preserve and share

https://modeldatarcn.github.io/

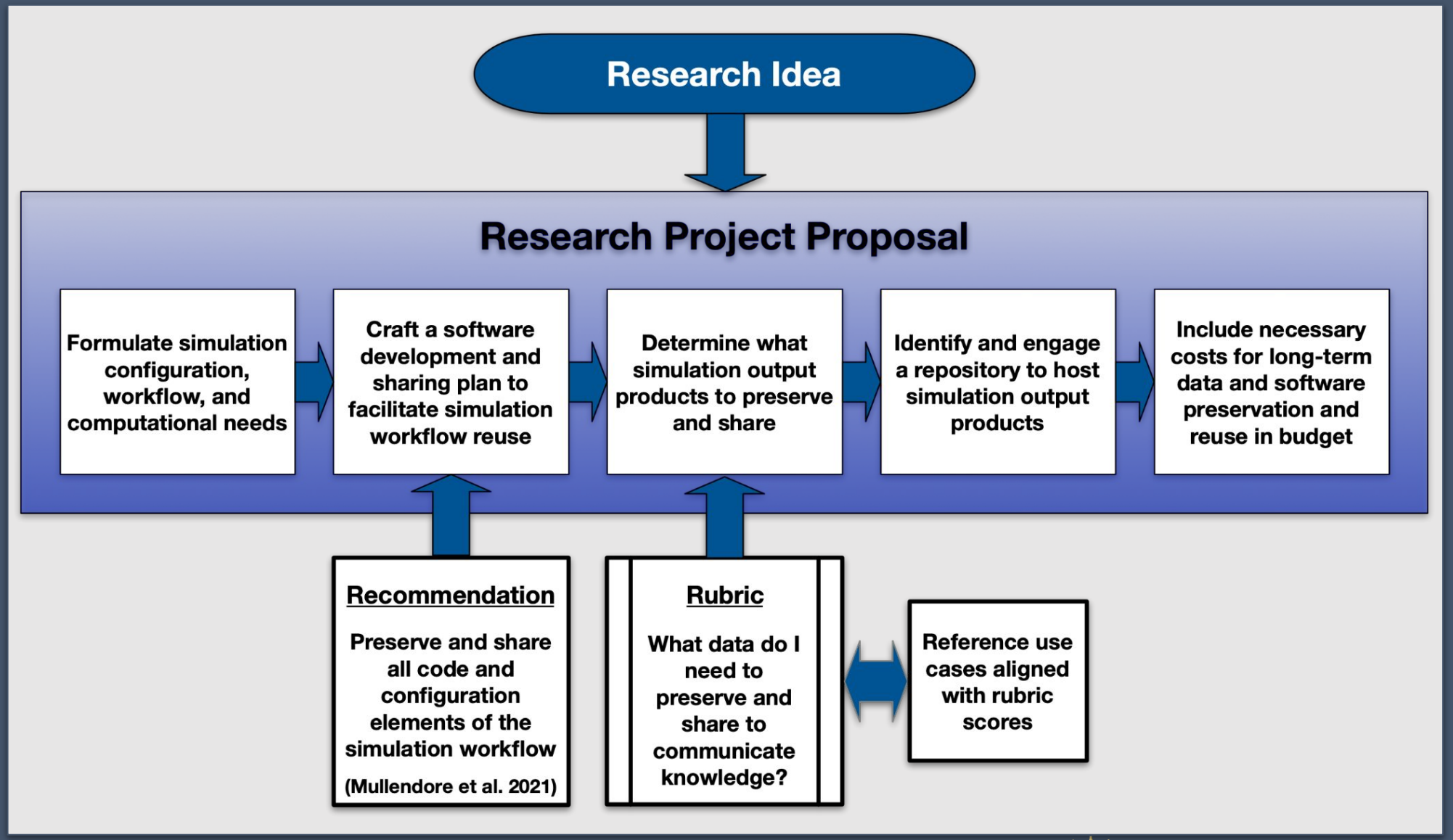
Rubric Purpose

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

When to Use: During project formulation phase

https://modeldatarcn.github.io/

Project findings -When to employ the RCN project guidance?



https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Community Commitment (Section Total Score: Min=3, Max=18)

- Is it anticipated that your simulation workflow outputs will have broad community impact and downstream reuse? No -Preserve less, Yes -Preserve More

Repository Data Accessibility (Section Total Score: Min=2, Max =12)

- Does the trusted community repository that you plan on archiving your data in provide adequate data access capabilities for the volume of data that you plan on depositing?

Simulation Workflow Accessibility (Section Total Score: Min=4, Max=12)

- Would it be straightforward for others in your academic discipline to rerun your simulation model run workflow steps?

Simulation Post Processing Workflow Accessibility (Section Total Score: Min=3, Max=9)

- Would it be straightforward for others in your academic discipline to rerun your simulation post processing workflow steps?

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Community Commitment (Section Total Score: Min=3, Max=18)

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

Repository Data Accessibility (Section Total Score: Min=2, Max =12)

- Does the trusted community repository that you plan on archiving your data in provide adequate data access capabilities for the volume of data that you plan on depositing?
  No -Preserve Less, Yes -Preserve More

Simulation Workflow Accessibility (Section Total Score: Min=4, Max=12)

- Would it be straightforward for others in your academic discipline to rerun your simulation model run workflow steps?

Simulation Post Processing Workflow Accessibility (Section Total Score: Min=3, Max=9)

- Would it be straightforward for others in your academic discipline to rerun your simulation post processing workflow steps?

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Community Commitment (Section Total Score: Min=3, Max=18)

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

Repository Data Accessibility (Section Total Score: Min=2, Max =12)

- Does the trusted community repository that you plan on archiving your data in provide adequate data access capabilities for the volume of data that you plan on depositing?

Simulation Workflow Accessibility (Section Total Score: Min=4, Max=12)

- Would it be straightforward for others in your academic discipline to rerun your simulation model run workflow steps? Yes -Preserve Less, No -Preserve More

Simulation Post Processing Workflow Accessibility (Section Total Score: Min=3, Max=9)

- Would it be straightforward for others in your academic discipline to rerun your simulation post processing workflow steps?

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Community Commitment (Section Total Score: Min=3, Max=18)

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

Repository Data Accessibility (Section Total Score: Min=2, Max =12)

- Does the trusted community repository that you plan on archiving your data in provide adequate data access capabilities for the volume of data that you plan on depositing?

Simulation Workflow Accessibility (Section Total Score: Min=4, Max=12)

- Would it be straightforward for others in your academic discipline to rerun your simulation model run workflow steps?

Simulation Post Processing Workflow Accessibility (Section Total Score: Min=3, Max=9)

- Would it be straightforward for others in your academic discipline to rerun your simulation post processing workflow steps? Yes -Preserve Less, No -Preserve More

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Research Workflow Output Accessibility (Section Total Score: Min=1, Max=6)

- Would it be straightforward for others across academic disciplines to use your simulation workflow outputs? No -Preserve Less, Yes -Preserve More,

Research Feature Replicability (Section Total Score: Min=1, Max=9)

- Would it be feasible for others in your academic discipline to replicate a feature generated through your simulation within an acceptable range of error?

Cost of Running Simulation Workflow (Section Total Score: Min=2, Max=12)

- What is the cost to product your simulation workflow outputs?

Repository Data Management Services Cost (Section Total Score: Min=1, Max=12)

- What is the cost to archive your output in a trusted community repository to preserve and provide access to your simulation workflow outputs for a minimum period of time?

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Research Workflow Output Accessibility (Section Total Score: Min=1, Max=6)

- Would it be straightforward for others across academic disciplines to use your simulation workflow outputs?

Research Feature Replicability (Section Total Score: Min=1, Max=9)

- Would it be feasible for others in your academic discipline to replicate a feature generated through your simulation within an acceptable range of error?
  Yes -Preserve Less, No -Preserve more

Cost of Running Simulation Workflow (Section Total Score: Min=2, Max=12)

- What is the cost to product your simulation workflow outputs?

Repository Data Management Services Cost (Section Total Score: Min=1, Max=12)

- What is the cost to archive your output in a trusted community repository to preserve and provide access to your simulation workflow outputs for a minimum period of time?

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Research Workflow Output Accessibility (Section Total Score: Min=1, Max=6)

- Would it be straightforward for others across academic disciplines to use your simulation workflow outputs? Yes -Preserve More, No -Preserve Less

Research Feature Replicability (Section Total Score: Min=1, Max=9)

- Would it be feasible for others in your academic discipline to replicate a feature generated through your simulation within an acceptable range of error?

Cost of Running Simulation Workflow (Section Total Score: Min=2, Max=12)

- What is the cost to product your simulation workflow outputs?
  Relatively Cheap and Straightforward -Preserve Less, Expensive and Complex -Preserve More

Repository Data Management Services Cost (Section Total Score: Min=1, Max=12)

- What is the cost to archive your output in a trusted community repository to preserve and provide access to your simulation workflow outputs for a minimum period of time?

https://modeldatarcn.github.io/

Rubric -Simulation Descriptor Themes

Research Workflow Output Accessibility (Section Total Score: Min=1, Max=6)

- Would it be straightforward for others across academic disciplines to use your simulation workflow outputs? Yes -Preserve More, No -Preserve Less

Research Feature Replicability (Section Total Score: Min=1, Max=9)

- Would it be feasible for others in your academic discipline to replicate a feature generated through your simulation within an acceptable range of error?

Cost of Running Simulation Workflow (Section Total Score: Min=2, Max=12)

- What is the cost to product your simulation workflow outputs?

Repository Data Management Services Cost (Section Total Score: Min=1, Max=12)

- What is the cost to archive your output in a trusted community repository to preserve and provide access to your simulation workflow outputs for a minimum period of time? Expensive relative to budget -Preserve less, Cheap relative to budget -Preserve More

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 Descriptor(s)**					

https://modeldatarcn.github.io/

Rubric Structure

Simulation Descriptor Theme						
Big Picture Question	**Simulation Descriptor(s)**	**Simulation Descriptor Classes**			**Raw Score**	**Weighted Score**
		Class 1 **Preserve few outputs**	Class 2 **Preserve selected outputs**	Class 3 **Preserve most outputs**	1, 2 or 3	Depends on weighting

https://modeldatarcn.github.io/

Cost

Is it more cost effective to rerun a full simulation workflow or preserve model output products in a trusted repository?

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x2)
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 Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x2)
What is the cost to produce your simulation workflow outputs?	**Computational Cost of Running the Simulation Workflow**					
	Human Resource cost of producing the simulation workflow					

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x2)
What is the cost to produce your simulation workflow outputs?	**Computational Cost of Running the Simulation Workflow**	Small computational cost, no special platform needs			1	1
	Human Resource cost of producing the simulation workflow	Trivial effort required to replicate simulation for most end users			1	1

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x2)
What is the cost to produce your simulation workflow outputs?	**Computational Cost of Running the Simulation Workflow**	Small computational cost, no special platform needs	Moderate computational cost, easy access to needed platforms		1 or 2	1 or 4
	Human Resource cost of producing the simulation workflow	Trivial effort required to replicate simulation for most end users			1 or 2	1 or 4

https://modeldatarcn.github.io/

Section Theme: Cost of Running Simulation Workflow (Total Score: Min=2, Max=12)

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x2)
What is the cost to produce your simulation workflow outputs?	**Computational Cost of Running the Simulation Workflow**	Small computational cost, no special platform needs	Moderate computational cost, easy access to needed platforms	High computational cost. Need specialized compute capability...	1, 2 or 3	1, 4 or 6
	Human Resource cost of producing the simulation workflow	Trivial effort required to replicate simulation for most end users		Significant time & expertise required to replicate simulation...	1, 2 or 3	1, 4 or 6

https://modeldatarcn.github.io/

Section Theme: Repository Data Management Services Cost

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x3)
What is the cost for you to archive the output in a trusted community repository..?						

https://modeldatarcn.github.io/

Section Theme: Repository Data Management Services Cost

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x3)
What is the cost for you to archive the output in a trusted community repository..?	**Repository Supported Data Curation Cost**					

https://modeldatarcn.github.io/

Section Theme: Repository Data Management Services Cost

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x3)
What is the cost for you to archive the output in a trusted community repository..?	**Repository Supported Data Curation Cost**	Community repository data curation expenses are prohibitive due to large volume of the expected model outputs.			1	1

https://modeldatarcn.github.io/

Section Theme: Repository Data Management Services Cost

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x3)
What is the cost for you to archive the output in a trusted community repository..?	**Repository Supported Data Curation Cost**	Community repository data curation expenses are prohibitive due to large volume of the expected model outputs.	Moderately expensive		1, 2	1, 6

https://modeldatarcn.github.io/

Section Theme: Repository Data Management Services Cost (Total Score: Min=1, Max=12)

Big Picture Question	Simulation Descriptor(s)	Simulation Descriptor Classes			Scoring	
		Class 1 **Preserve few**	Class 2 **Preserve selected**	Class 3 **Preserve most**	Raw Score	Weighted Score (x3)
What is the cost for you to archive the output in a trusted community repository..?	**Repository Supported Data Curation Cost**	Community repository data curation expenses are prohibitive due to large volume of the expected model outputs.	Moderately expensive	Would be inexpensive to curate the complete simulation workflow output for a minimum number of years in a community repository.	1,2 or 3	1, 6 or 12

https://modeldatarcn.github.io/

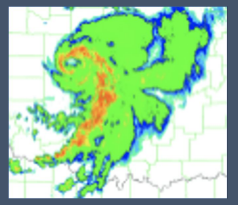
Rubric -Total Score of Descriptor Section Themes

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/

Preserve Few Simulation Workflow Outputs (Score < 48)

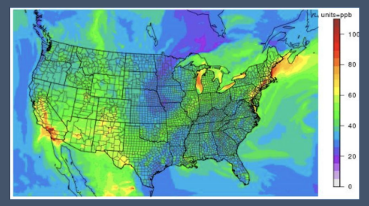


Idealized Process Study – Goal is knowledge production, most value in model configuration and codes

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

Ensemble Forecast Experiment – Important environmental fields are saved in the form of "summary files", which are a fraction of the raw output

Preserve Majority of Simulation Workflow Outputs (72 <= Score)



Modeled Ammonia Emission Profiles -Goal is data production for downstream reuse

https://modeldatarcn.github.io/

Additional RCN Findings

- **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/

EarthCube
Transforming Geosciences Research

Additional RCN Findings

- **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/

EarthCube
Transforming Geosciences Research

Additional RCN Findings

- **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/

EarthCube
Transforming Geosciences Research

Additional RCN Findings

- **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/

EarthCube
Transforming Geosciences Research

https://modeldatarcn.github.io

AMS Data and Software Policy Guidelines for AMS Publications
https://ametsoc.org/PubsDataPolicy

AGU Guidelines for Research Primarily Based on Numerical Models or Theory
https://data.agu.org/resources/agu-data-software-sharing-guidance#guidelines

Example Interactive Rubric:
https://modeldatarcn.github.io/rubrics-worksheets/rubric-example.html

Questions? schuster@ucar.edu, mayernik@ucar.edu

https://modeldatarcn.github.io/