General Workshop Notes:

08/05/20

Morning pre-breakout

- Knowledge production vs data production discussion
    - Article https://esajournals.onlinelibrary.wiley.com/doi/10.1002/ecs2.3191
    - Current rubric is targeted at the knowledge production use case
    - Examples of data production
        - CMIP
        - Operational forecast models
        - Reanalysis
    - Examples of knowledge production
        - Created to support a specific paper
        - Data produced through "data production" projects can be used for knowledge production
    - Decision tree above rubric
        - There will need to be up front questions to address this type of use case prior to diving into the rubric
        - There may be other similar use cases that don't need to use the rubric, will need to develop similar "up front" questions for those.
            - E.g. is storage not an issue (< 10 GB of output expected)
- Consider the reproducibility side of this
    - How can reproducibility be achieved on the knowledge side?
        - What needs to be preserved to support reproducibility of the research findings?
        -
- Google chat notes:
    - Cathy Smith - NOAA Affiliate9:18 AM
    - The question of who 'owns' data in repositories is an issue. Particularly if there are updates/errors/changes.
    - Jonathan Petters9:21 AM
    - And those terms are repository dependent!
    - Shelley Stall9:21 AM
    - Happy to help get the word out to the larger publisher community on your recommended guidelines on what should be preserved/cited. Working with Mike Friedman and AMS journals of course.

Breakout session followup

- **Breakout 1 notes**

- **Breakout 2 notes**
    - what is the intended use? (at proposal stage)
- **Breakout 3 notes**
    - Knowledge production
        - Need to aggressively cull the simulation output
        - How well workflows are being documented
        - Save workflow used to support knowledge production
        - Not all data are typically need to support scientific reproducibility
    - Data production
        - Very few projects fit into the “data production” side
        - Need a coordinated approach to archiving and curation
            - Needs investment in personnel and resources in general
            - Coordinated -not project to project
                - E.g. NSF Excede model
            - Could private industry play a role here?
                - E.g. Cloud “Open Data” programs
                - Could they even work to compute/produce the data?
                - Need paying end users to pay for cloud compute colocated with the data. These end users likely exist in the risk management community
        - Need better communication to end users on data management resource constraints.
            - Can't save everything

- **Breakout 4 notes**
    - Strawman starting point
        - Model data not useful after 5 years?
        - Older data are smaller, so storage isn’t an issue for those legacy datasets
    - Consistent data access metrics across repositories can be challenging
    - Use cases for legacy model data
        - Comparison studies
        - Educational purposes -simple use cases for students to start out with
        - Capability to fix errors, or highlight errors from past runs

- Versioning
    - What to keep and how long to keep
- Where are data saved/preserved?
    - Cloud vs other services?
    - NOAA datasets in the cloud have seen large increases in access
    - Barriers to access can be mental
    - Need programmatic access capabilities
    - Would there be benefits to a single resource for model output archival?
    - Community resources
        - EG hydroshare
- Who pays
    - NSF funded repositories for some disciplines
    - General purpose repositories for smaller datasets -free
        - Zenodo, Figshare
    - Could services -good for large data that are going to be used a lot (data production datasets)
- Communication to stakeholders
    - Be clear on up front costs
    - Communicate as a community -larger impact
        - Illustrate consensus or agreement within a community
- Policy -Journal of water resources and management
    - Take more of a carrot vs the stick approach
        - Reward publications that try to make things more reproducible
            - Provide great examples for others to aspire to
        - Reproducible publications are on the spectrum vs being on the binary
            - Incremental improvement vs a high bar
- General comment
    - Where might we want this to get to. What do we aspire to?
        - If data are well documented and accessible, would there be more data reuse?

How to move forward

- Rubric and companion document to be further developed

    - Companion document -findings and recommendations on selected challenges
- Ask for the workshop community to start using and testing the rubric, and provide feedback on your experiences.
- RCN is holding a Town hall at the AMS annual meeting
- Participants are invited to promote the project through any of their discipline specific community resources
- Important next steps from participants
    - Engage publishers on what policies have been developed thus far
    - Need more coordination at NSF on Big Data archiving and funding on how to do this
        - Review findings and challenges from the workshop, and how to address these?
    - Tailored rubrics by communities or departments within an entity?
        - GitHub repo for modified versions of the rubric?
    - Add a decision tree above the rubric?
        - Yes. Especially data production vs knowledge production
    - Communication -proposal costs can be uncertain.
        - How much data volume will end up needing to be preserved in a CTS repository?
            - Is it reasonable to ask this?
            - Can you ask for more resources later as you can do with computing?
            - NCAR asks for an “order of magnitude” estimate at the project proposal phase
        - Clark could help engage “monthly weather review” -AMS publishing on this topic
        - Matt -engage Mike Friedman
    - What participants would be willing to participate in follow on workshops or proposals?
        - A number of participants are responding “Yes” on the chat
    - Thanks to everyone
    - Email us with ideas

Chat transcript from 11AM - 12PM MDT

- Susan Borda11:24 AM
- Gretchen/group 3 - "need a coordinated approach to archiving and curation", glad to hear it!
- Gary Strand11:25 AM
- Thanks, Gretchen - that was great!
- Glen Romine11:25 AM
- perfect, thanks!
- Amy McVey11:39 AM
- I plan to use the Rubric and make one specific to my department so we can all be organized and on the same page.
- Susan Borda11:42 AM
- I'm putting together a list of pre-deposit questions based on the rubric for data depositors/researchers.
- Ruth Petrie11:44 AM
- I'll be suggesting that we use the Rubric in the UK to help CEDA as data curators help the researchers decide what data to archive. I'll suggest this approach also goes into the funding proposals in the "data production box" and highlighting that it is ok to say it is a knowledge production project and only minimal data will be produced. I'll let you know how we get on. This will be extremely useful. Thanks!
- Susan Borda11:46 AM
- Yes I would be interested.
- Kevin Tyle11:46 AM
- I'd be happy to participate in further workshops
- Maegen Simmonds11:46 AM
- I plan to bring these discussions back to the ESS-DIVE data repository to see if we can use something like it in guidelines I'm developing for model data archiving, as well as in a paper on this.
- Also happy to participate later on
- Dan Tyndall11:47 AM
- i am willing to participate in another workshop
- Katelyn Barber11:47 AM
- ditto

- Jamie Wolff11:48 AM
- Happy to remain involved.
- Amy McVey11:48 AM
- I'm interested in helping more.
- Jonathan Petters11:48 AM
- Sure, happy to be involved
- Gary Strand11:48 AM
- I'll be happy to participate!
- Maegen Simmonds11:48 AM
- will do!
- Leslie Hsu11:49 AM
- @Maegen I'd like to connect to share notes on model data archiving, as USGS is also exploring this now, I will find you online....
- Maegen Simmonds11:49 AM
- yes great, @leslie! mbsimmonds@lbl.gov
- Kevin Tyle11:50 AM
- Thanks to you, Gretchen, Doug, and Matt for organizing and running a stimulating meeting! Hope to see folks again in person sometime sooner than later!
- Matthew Mayernik11:50 AM
- Project webpage where the outcomes will be posted: https://modeldatarcn.github.io/
- Leslie Hsu11:50 AM
- Yes, thank you, organizers!
- Maegen Simmonds11:51 AM
- Very happy with how this went - very relevant, important stuff, and very engaging. Thank you!
- Ruth Petrie11:51 AM
- Thanks for a great meeting, learned lots!
- Tiffany Vance - NOAA Federal11:51 AM
- Thank you to all the organizers. incredibly smooth transition to virtual Let's hope for in-person in 2021.
- Gary Strand11:51 AM
- Thanks!
- Ted Mansell - NOAA Federal11:51 AM

- Thank you all!
- Dan Tyndall11:51 AM
- bye everyone

Breakout leaders follow up

- Elisa
    - Small group -4 participants
    - Mentoring came up -need to capture the importance of this at early career
        - Supports significant culture change
    - Add mentoring educational component as part of future workshop
    - Everyone replied “save everything” at conclusion of project in group 1
- Jared
    - Psychology plays a role
        - Keep all data adds stability to very dynamic lives
        -
- Gretchen
    - Culling data is learned through experience
    - Early career can lead the charge

08/04/20

Morning pre-breakout

- Need to score components of complex workflows separately as each component may score differently
- Discuss qualitative vs quantitative use cases for the rubric
- Rubric is very subjective, and will depend on who is using it. The perspective of the rubric use will vary.
    - EG Discipline specific/expert users might think WRF is easy to use, while users from other disciplines might not.
    - Entry level users might answer differently than senior level users
- **Gretchen -right now we communicate to discipline specific (peer group) users**
    - Susan Borda -Downstream uses/users is what we are hoping for in the academic library repository community.
    - To support downstream users, there would generally be a need to “save more data”
    - EG climate data can be used many ways by downstream users, including regional modelers.
    - Gretchen, does this get into “data production” vs knowledge production?
        - Save workflows vs save all of the output?

- Glen -due to resource constraints, can save only data to answer specific science questions
    - Datasets made to be shared such as reanalysis don't fall into this category
- Clark
    - How frequently do people reach out to others for data requests?
    - Ratio of computing resources to where data are saved needs to be considered
    - What is the extent of the target user audience?
        - Share code for niche projects?
        - Save/share data for broader use projects?
- Gretchen
    - DATA production -Save much more
        - Reanalysis -Useful for a large number of studies
        - CMIP -useful for a large number of studies
    - Topic of this workshop -Knowledge production
        - Save for portions of code, data, and workflow for targeted audiences
- Kevin T
    - Guidance from federal agencies and publishers for reproducibility can be a challenge
        - Do we feel comfortable enough to not store everything and fulfill those requirements?
        - Gretchen -Hopefully this rubric will inform those requirements (cite this rubric in data availability statement or project proposal)
    - Cindy
        - If a dataset falls into "Data Production -high community resource dataset"
            - This might need to be weighted heavier
            - See Rows 17 and 18 in v1.1 of the matrix
- Save more data for complicated workflows? -touch to reproduce. Related to weighting of descriptors.

- Chat comments from AM general session
    - Susan Borda9:12 AM
    - Downstream uses/users is what we are hoping for in the academic library repository community.
    - Susan Borda9:15 AM
    - documentation, documentation, documentation
    - Ted Mansell - NOAA Federal9:18 AM
    - I like the idea of saving workflow, which would be cheap and important for reproducibility and for when a model is updated/changed (and hopefully doesn't break the workflow!)
    - Susan Borda9:20 AM

- I'd like to know what features repositories should have to make the data most useful to researchers.
- Glen Romine9:26 AM
- consider having data classes - where for some data classes, the score doesn't really matter. Could be community definition of class, instead of the data generator self-defining to reduce bias.
- Kevin Tyle9:30 AM
- Obviously, papers and presentations are done subsequent to one filling out the rubric, but for reproducibility's sake, including the code that produces figures would be good to commit to saving.
- Richard Neale9:32 AM
- I'm not sure if we've talked about this, but we have to consider the cost of rerunning simulations for example. A observational campaign has a high cost of 'rerunning'. A single column run has a low cost.

Breakout followup

- **Breakout 1 notes.** **(theme: save less data)**
  - Code documentation and version control to track changes is essential
    - Especially where niche changes are made.
      - Minimum of diff b/w old and new code, and comments as to why and how these changes were made.
      - Comment: why and how is its provenance.
    - Save config (Namelists)
  - Little to no raw output needs to be preserved for the general community
  - Non-linear case brought of issues of feature reproducibility
    - Compiler and processor dependent
    - Comment: Feature reproducibility with respect to compiler issues can be taken care of by using containers and tracking dependencies.
  - Processed output
    - Specific to research study
      - E.g. tropical cyclone intensity and land surface parameters
    - Community -validate that research makes sense in a broader perspective
  - Time horizon -keep data, keep codes, maintain DOIs?
    - Group thought a 5-year time horizon is a good place to start in making these choices
    - 5-years after publication or 5-years after creation? (After publication might be more relevant...but an open question)
  - Chat

        - anu Malik11:25 AM
        - Feature reproducibility with respect to compiler issues can be taken care of by using containers and tracking dependencies.
        - Dan Tyndall11:27 AM
        - tanu, will a docker image produced today be compatible with computer hardware and docker software 5-10 years from now?
        -
- **Breakout 2 notes** **(Theme: leans towards save less data)**
    - Scientific value of reproducibility?
    - Comment: I think reproducibility is the cornerstone of science so certainly there is value. Most science builds on previous results being reproducible. The problem is that a lot of science is in transient stage, and it is not clear if transient science results should be maintained in a reproducible form (which requires lots of resources). In that sense, it becomes an optimization/decision problem--save more or save less.
        - Reproducibility is Field specific
            - Are the results scientifically feasible in a discipline specific community
        - https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html
- **Breakout 3 notes** **(Theme: leans towards save more data)**
    - 3 use cases
        - 2 on the WRF model + study for Amazon
    - Keep config/namelists
    - Input data
        - Save processing codes and documentation for processing codes
        - Point to input data
            - Can be challenging when there is not a repository for input data (e.g. GEFS)
    - Amazon study
        - Save data every 10 minutes over region of interest
        - Larger area, save data every hour
    - 3d ocean model and set of GCM simulations
    - Thorough documentation is essential
        - Includes codes for generating plots for papers, etc..
    - Save post processed output
        - Much smaller than raw WRF output
- **Breakout 4 notes** **(Theme: Save most data)**
    - WRF based project
        - Purpose of the project has a “data production” component driven by funder requirements that leads to “Save most data”
    - Land Use project based on carbon footprint

        - Small output volumes
        - Audience: Its essential to preserve all outputs for decision makers who use this data
        - Audience wouldn't know how to rerun the model
    - CMIP -global scale -data production
        - Large downstream user community
        - Still need to have discussions on how to selectively reduce data volume
            - Compression choices
            - Save selected parameters/levels
    - Takeaways
        - Development runs vs production runs
            - Development
                - Can include 10s-100s of runs that produce data with limited value
        - Comment: I find this distinction quite valuable. In the development runs the emphasis can be on code and inputs, instead of outputs, where as for production runs the outputs become increasingly more important.
        - Reasons for keeping lots of data
            - Computational barriers: Difficulty/cost in regenerating outputs for yourself or outside users
            - Target user community: Users expect specific outputs
            - Model code is not sharable (IP or nat security)
            - Users rely on HPC centers to produce these outputs because they have the resources to do so.

    - Chat notes
        - Dan Tyndall11:27 AM
        - tanu, will a docker image produced today be compatible with computer hardware and docker software 5-10 years from now?
        - I don't have a lot of familiarity with docker
        - Tanu Malik11:27 AM
        - Docker provides no such guarantees.
        - Dan Tyndall11:28 AM
        - MPI reduce :)
        - Tanu Malik11:32 AM
        - Intermediate datasets can have value for recomputation/reproducibility.
        - Cindy Bruyere11:33 AM
        - Agree Tanu - which made me think that I will come up with different results if I split my workflow before applying the Rubric.

- Tanu Malik11:34 AM
- I think asking for lifetime of containerization or container images is a wrong question but what are the fundamental system primitives that can be preserved over a longer span. This currently is an active research issue.
- Thanks for the summary Doug.
- Tanu Malik11:35 AM
- I agree Cindy.
- Susan Borda11:41 AM
- nice summary Adam
- Tiffany Vance - NOAA Federal11:41 AM
- Good summary Adam
- Katelyn Barber11:42 AM
- Yes, post processed
- Tanu Malik11:42 AM
- If input data is not preserved then how can the results be reproduced?
- Is the source code for regenerating input data available?
- Dan Tyndall11:43 AM
- well, that goes back to the question of whether or not the authors need to store the input data if it is gotten from another repository
- You11:43 AM
- Agree. Who's responsible for preserving input data? We're thinking of operational model output produced by NOAA/NCEP
- Tanu Malik11:45 AM
- Yes, here the human part is crucial---without input data being preserved, reproducibility downstream can be broken
- Susan Borda11:49 AM
- In a general repository like mine, I would save the input data with the rest of the data package but perhaps keep it private/hidden from public then share it if the "canonical" version of the input data is no longer available. This way you have a snap-shot of the input data if it should change.
- Gary Strand11:50 AM
- Great summary, Matt. Thanks!
- Tanu Malik11:52 AM

- Susan, Yes snapshots are important for exact reproducibility but even if approximate input is available, I have heard users consider that as very useful.
- I meant "I have found" instead of "I have heard"
- Glen Romine11:58 AM
- Accessibility (e.g., mechanisms of access) of the data is worth discussing as well.
- Jonathan Petters11:58 AM
- Agreed Matt....understanding how fast different met/climate communities move and how long these things might have value would be really useful

Breakout Leaders follow-up discussion

- Gretchen -Original Day 3 plan -cross validation
    - What topics would you like to tackle tomorrow?
        - Where does data go?
        - How long does it need to be saved?
        - Who pays for it?
        - Data production choices
            - Downstream users (build upon the research of others) vs reproducibility (scientific integrity/trust)?
        - Who's responsible for data once the student moves on?
        - Questions?
            - What to save?
            - Where does it go?
            - How long is it there?
            - Who pays for it?
            - What type of access is provided?
            - Is it practical to save based on available resources?
            - How long to save data or code from the science perspective?
                - When does data lose science value?
- Narrative that goes with rubric will need to capture nuanced issues.

08/03/20

Morning Plenary

- This document wasn't available for the 8/3/20 morning plenary discussion. Find plenary presentation slides at:
    - https://modeldatarcn.github.io/workshop2/presentations/presentations.html
- Google Hangout Chat during the plenary session
    - Comments related to Gretchen's use case
        - Tanu Malik: What programming languages is the model coded in?
        - Ted Mansell - NOAA Federal: Could store at least a base model run for comparison (e.g., results vary slightly by compiler and optimization)
        - Tanu Malik: When you say you made some modifications---were they source code modifications or just input parameter modifications?
        - Jeff de La Beaujardiere: Ted's point is good -- maybe when we don't save the output, we should at least save either a representative output file, or statistics about the output (extrema, means, standard deviations, maybe even skew and kurtosis), or some visualizations.
        - Gretchen Mullendore: Excellent point, Jeff (and Ted). I was planning that a whole suite of output plots/stats would be saved, but I didn't state that explicitly. I should not have stated "no output" as derived output was planned. Great thing to keep in mind as we move into days 2 and 3 this week.
        - Susan Borda: Gretchen, what sort documentation would you expect to include with your data?
        - Cathy Smith - NOAA Affiliate: How should model changes be documented?
        - Gretchen Mullendore: Susan and Cathy: yes, documentation! A hugely important part of a save to repository. Something we need to work through as well. We all know that the publication alone is often not sufficient documentation. I welcome everyone's ideas on basic standards for documentation.
    - Comments on Matt's use case
        - Mimi Hughes: NOAA Federal: We're not doing tests of this kind today, correct? I meant specifically CMIP type :)
        - Susan Borda: Matt, if you don't save "everything" what expressly would you NOT save?
        - Matt Mayernik: Thanks Susan. That question is what we hope to discuss in detail during tomorrow's breakouts.
    - Comments on Adam's use case
        - Dan Tyndall: When we are "saving all data", should we be duplicating other repositories? For example, prepbufr data is hosted on NCAR RDA. If I archive my other data elsewhere, but rely on RDA for prepbufr, and that disappears, I may not be able to rerun my case
        - Glen Romine: does it really make sense to aggregate/average the scores? Some aspects seem like they should have a bigger weight than

others. Comparing the CMIP and WOF data sets, one might expect larger separation in scores.
        - Kevin Tyle: I was also wondering, as Dan does above, about how to take advantage of well-established community repositories such as RDA. Haven't considered the scenario of a repository such as that disappearing at some point, but does need to be considered!
        - Gretchen Mullendore: Glen: One of the questions we had as well. Should we weight different model descriptors differently? Or group certain descriptors together? I hope that as we get "what we save" examples fleshed out, we will see areas that need to refinement.
        - Jeff de La Beaujardiere: Should the rubric have a row asking about long-term usefulness of the model output? For example, something like Warn-on-Forecast sounds extremely useful for the hours covered, but by the next day will presumably be of much less relevance. (I realize Adam said in this case not to save the model outputs, but there may be other use cases where long-term usefulness is perhaps low.)
        - Glen Romine: the RDA, I don't think, has prepbufr data used in the WoF (HRRR) which is hourly generated.
    - Comments on Laura's use case
        - Tanu Malik: Sorry for the dumb question (I am a computer scientist) ---What is the difference between integrated model and coupled models?
        - Jonathan Petters: Is the time involved in learning to use model like HyrdoFrame to recreate/modify model output included in the rubric under "Human Effort?"
        - Susan Borda: Laura, how often are you generating this data (13TB+) , annually or less often (or more often)?
        - Gary Strand: Coupled models consist of multiple component models that interact with each other (atmosphere, ocean, sea ice, etc. for global climate models); they are a kind of "integrated model", which is a generic term.
        - Tanu Malik: Thanks, Gary.
        - Dan Tyndall: I think the speaker makes a good point about the minimum being established. My use case involves a closed source model we generally don't release to the public. Sometimes we are evaluating confidential observatuons--the forecast data produced by the obs are ok to release, but not the obs themselves
    - Comments on Brooks' presentation
        - Kevin Tyle: Would Github be an example of an "appropriate repository" for the purposes of software curation?
        - Tanu Malik: To organizers: Will slides from presenters be available for reflection?
        - Jasmin John - NOAA Federal: How do you ensure citation of data?

        - Jasmin John - NOAA Federal: Thanks. I've come across papers that are not citing CMIP6 data even though it is required if using the data. Hence my question.
        - Brooks Hanson: thanks jasmin; good to know; if you have an example that would be good

Breakout followup

- This document wasn't available for the 8/3/20 post breakout discussion. Find individual breakout session notes below
    - **Breakout Group 1 notes**
    - **Breakout Group 2 notes**
    - **Breakout Group 3 notes**
    - **Breakout Group 4 notes**
    - **Breakout Group 5 notes**

Breakout leaders followup discussion

- Gretchen
    - Workflow elements need to be scored differently
        - Scores can be different depending on the various elements
    - Descriptors should be weighted by users
        - Depends on the individuals use case
        - This is to be used as a tool, not to provide an authoritative answer
- Breakout Day 2 -A few summary use cases for your scoring group.
    - Use case types.
    - Two or three use cases
        - Short description
        - What do you think should be saved for this use case?
            - Possible workflow elements: input, model code, model setup, raw output, processed output, pre- or post-processing code
            - Documentation
        - Why should this be saved?
    - Score discussion on different workflow elements?
- Documentation related:
    - Guide to writing README description files:
        - https://data.research.cornell.edu/content/readme