Compiled notes from Model Data RCN Workshop 1, May 5-7, 2020

- **Assumption: Rubric to be used to assist a researcher in determining what data or software should be deposited in a FAIR aligned repository to communicate knowledge.**

Compiled notes sorted according to the descriptor themes

- **Accessibility**
    - How does compression fit in? Involves some tradeoffs among categories, e.g. space vs usability.
    - Different users may need raw data vs post-processed data
        - Weather research - don't have freedom to run models, post process, then decide what to look for. Tend to decide what to look for before running models
        - Climate - save lots of data to allow exploratory studies by broad groups of people.
    - What do we mean by "saving"?
        - E.g. CanESM5 has 500 TB on ESGF, 5PB on local (non-public disk) and 30PB on tape. So there might also be "tiers" of how/where things are saved and how accessible they are.
        - When does clock start for recommended "save" term?
    - We need to keep an eye on whether the data are accessed or not we might want to remove data that are never access
    - If we save files in the form of single variables, it is easier to figure out whether it is used or not and to delete what is not used.
    - Taking things out of archives is really hard and storage is expensive
    - Archive things continuously or do it at the end of the project. Its hard at the end of a project to get people to archive things at the end
    - Datasets are not really easy to archive often and unless there is a workflow that goes with a dataset then its really hard for anyone to use it.
    - How can you do a publication that uses data and not support it somehow? Doing a data dump is not necessarily the answer. In some cases the research will really need to be substantiated with the collection. If we succeed how will we substantiate our findings.
    - Try to explain this as a cradle to grave way of dealing with data - or estate planning for your data legacy
    - Need to prove that your findings can be **substantiated** - The data and the codes are not necessarily what is the most important. Reviewers should be able to check your workflows.

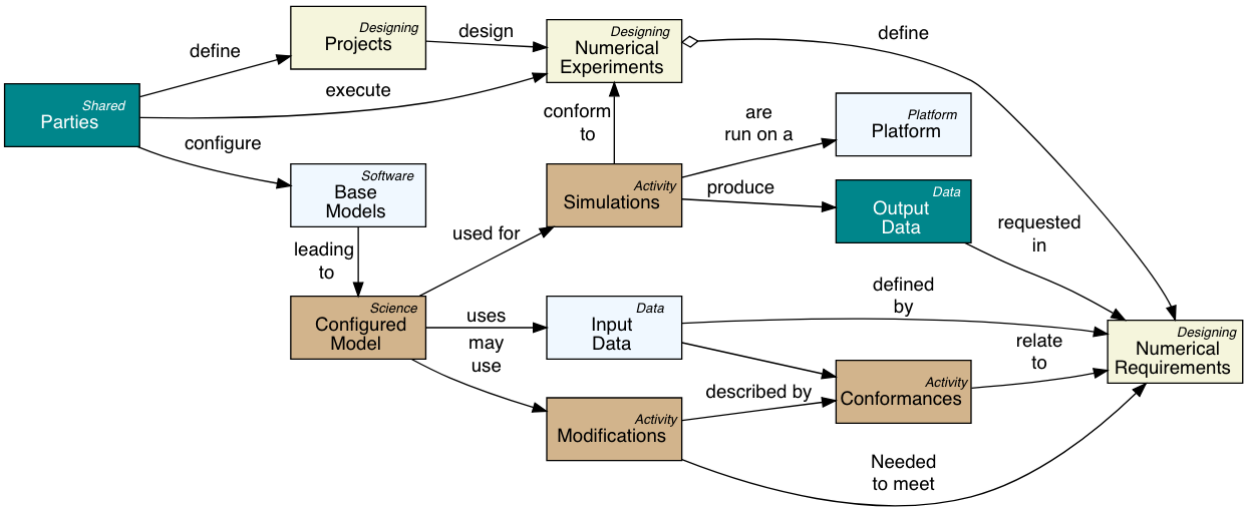
- How can we manage archives so that people can actually get what they want back out? Can they find what they need? Do they know what it is?
- There also needs to be a push so that people actually would want to use other peoples results rather than just running their own. -- need a way to incentivise this behavior
- We have a lot of redundancy of who is running what.
- People really want the global data. Once people start downscaling then it gets really specific and hard to share across groups
- Data is really not findable-- even if you can use someone else's data its really hard to find what you need.

- **Community Commitment**
  - Lines about "value to community" category are all oriented toward saving more output because they are assuming data being generated as part of other projects or data compilation efforts, e.g. large user communities already exist.
    - These rows largely more applicable to the climate community?
  - Need a broader community when deciding what to save -- for example with CMIP when making data that is meant for a broader community.

- **Cost**
  - Some models are so expensive to run that storage has to be planned ahead.
  - Could maybe sum cost sections up as is it more costly to save the output or re-run the models to regenerate the data.
    - "If I need this data again, is it cheaper to save or re-run"
    - Maybe not for data generated for a community.
    - Dependencies between how long it takes to run a model on your available computing systems.
  - If compute is co-located with the data, is that a reason to save more or less of the data? The more people can bring their compute to the data, the more they want to save their data.
  - Separate into two rows "computational cost" (CPU hours required) vs "specialized platform" -- these are orthogonal considerations, but are conflated in columns D-E-F of original sheet.
  - If a model is viewed that this is a dataset that is of value to downstream users that the PI needs to make the case to funders that resources are needed to really set this up.
  - Tornado simulation: running the model is so expensive that keeping the model output is cost effective.
    - Would have to go to the data to do computing, e.g. HPC center.
    - May change over time with cloud-based computing
    - Glen: mostly just save post-processed output, not raw output
      - May not want to obligate a researcher to save data for potential use by someone else.

- **Longevity**
  - Model longevity & versioning - depends on other factors.
  - Longevity of data might be higher than longevity of software, but both are limited.

- **Provenance**
  - Having a persistent identifier, to allow reference to a particular version. Perhaps as part of the citation.
  - Original writer of the dataset writes a paper that describes the data.
  - Are there metadata or documentation standards for netCDF data? I appreciate the self-describing nature of *.nc files, but is there a “data dictionary” or metadata standard that explains these variables?
    - NCEI recommends, but does not mandate, CF metadata conventions for netCDF model data
    - CEDA pretty much mandates the CF conventions.
  - We need to describe, separately, *something* about each of these entities. However, the amount of information needed is a function of need. For CMIP we want it all. For something else, we might want a lot less (actually even for CMIP, we probably ought to be more discriminating).



- Pie in the sky (list of CMiP6 models) and their descriptors:
  - https://search.es-doc.org/?project=cmip6&documentType=cim.2.designing.NumericalExperiment&client=esdoc-url-rewrite
  - (Change document type from experiment to model to see model descriptors)
  - The Value for model type is from a controlled vocabulary, which includes things like regional climate model etc, so one can (in a different UI) query the database for different types of model.
  - But in ESGF, if you want data, you can use the same type of controlled vocab, to find the data from particular types of model.

IPSL > IPSL-CM6A-LR :: Top Level	
Top Level Properties	
Top Level > Model Name	
Description	Name of coupled model
Value	IPSL-CM6A-LR
Top Level > Model Keywords	
Description	Keywords associated with coupled model
Value	IPSL, climate model, earth system model, LMDz atmospheric general circulation model, NEMO oceanic general circulation model, ORCHIDEE land surface model
Top Level > Model Overview	
Description	Top level overview of coupled model
Value	-- Awaiting modelling group input --
Top Level > Model Type	
Description	Model type
Value	GCM

- **Reproducibility**
  - Need to keep a realistic viewpoint for what people are being asked to do.
  - How to visualize the rubric? May be helpful to think about sliding scales, decision trees, matrices.
  - What is the starting point for use of the rubric?
  - Want consensus on results, not on any one person's result and their set up.
  - Last snapshot of data (data used to generate figure) is probably not the most useful data to be keeping around.
  - Like the concept of "Gray data" - data that is processed/filtered down somewhat, but not totally culled down to just the numbers behind the figure.
  - Need to be thinking about model-based science as involving workflows with a number of steps, each of which may have outputs that are inputs to the next step(s).
  - Consider deleting "bitwise reproducibility" criterion -- nearly impossible to achieve in all realistic situations, therefore should not be used in archiving decision. Maybe replace by notion of "numerical reproducibility" within some acceptable epsi
  - It seems as though a goal of reproducibility versus one of reusability may affect our discussions.
    - We can have model descriptors about levels of reproducibility AND reusability. Both may be useful.
  - For our CMIP6 / CanESM model development effort we used the terms "reproducible" to mean bitwise reproducibility and "repeatable", to mean we could reproduce the climate but not the bit-pattern (i.e. experiment reproducibility). For

CMIP6 we dumped about 30Pb to tape over ~12 months. We found it faster to rerun the model (bitwise) than to load data from tape in many instances - so saving the data is not always the best answer. As a modelling centre, this bitwise identity is practically very useful for e.g. hole-filling, while “repeatability” might well be sufficient for resulting scientific studies. For our CMIP6 NetCDF data we also include our git version control commits (SHA1 checksums) for the model source code, model configuration repository, and NetCDF conversion processing directly into the NetCDF metadata. Using these anyone can trace exactly which code produced a given file. The model will not run a production experiment if someone alters the code without committing the result. With respect to computational environments changing, there is a role for containerization in making runs at least repeatable, despite changing hardware.

- Distinction between reproducibility of simulation vs reproducibility of the experiment is very important
- Reproducibility categories - question might be whether authors have checked whether these matter (feature, statistical, bitwise)
    - Ensembles are important to the “feature” and “statistical” categories, but not “bitwise”

- **Value Judgement**
    - Value to community may only become apparent in retrospect.
    - Highly idealized simulations (or suites of idealized simulations) could be equally as valuable to save as simulations with as much realism as possible. For example, a suite of LES of convective boundary layers spanning a range of stabilities could be highly valuable as a database for developing boundary layer parameterizations for NWP, understanding basic physics, etc. May want to consider when developing use cases.
    - The value of saved data depends on other factors: how easily can the data be accessed, how much reduction/processing can be done near the data vs. after download, how well the data is described, etc. Deciding what data to save is a decision that probably can’t be made in isolation
    - Archived/Legacy datasets versus new datasets. Scientific value more applicable to (and easier to assess) what to preserve rather than what to save.
    - Consider value to disciplinary community but also external community (other disciplines, industry, public, policy-makers, education, etc)
    - Publications that get star rated to reflect if they have the data, workflows, and code available. And DOI’s for code, data, and workflows.
    - Archive enough to evaluate the work scientifically
    - Development runs more fluid, final runs more constrained. Modelers need to clean up workflow before they run - what are you trying to accomplish with an experiment/simulation
    - Value proposition of archiving and being able to compute on the data. Value proposition of archiving vs re-running simulations.

- Does the data provide any value to support an idea? -idealized simulations. Is the methodology enough?

- **Versioning**
  - Are there scenarios when you may need to take into account the "validated" version, not just the "standard" version of a model.
  - Model Versions and Longevity: Should distinguish between “base model code”, and “configurations” and “validated configurations” … the latter tend to be named, and you can say something about how “good” that model (configuration) is … which es-doc would describe using a quality_report document and/or a lineage document.
  - As we have incremental improvements -- do we deprecate the older simulations?

- **General comments/questions about the rubric**
  - What about issues of inequality nationally? How can we have a tiered structure of participation, but we also want smaller scale contributions.
  - How do embargos fit into the rubric?
  - We could think of a “descriptor” as a document, owned by someone, which describes a key part of the modelling workflow. Sometimes one person might own them all, and they might all be aggregated together into one larger document. In other cases, different people create them, and they exist in different places, and they are linked.
  - Is the rubric intended for both existing/legacy and future datasets?
  - Consider using a term such as "Modeling Project Descriptor" rather than "Model Descriptor" because it's not about the model code itself.
  - Not "raw data" but "model output"
  - Thinking about constraints at beginning of project to be cognizant of trade offs, e.g. if not resources to keep all output, emphasize having a well documented workflow, code, etc.
  - In general, suggest having as few Descriptors as possible to cover the landscape and make the rubric easier to fill out and evaluate. Could perhaps combine:
    - "Platform/System Availability" and "where/how was this run?" Mention cloud in platform rubric. Also combine Dependencies and Environments.
    - "Model/Code Availability" with "Model Re-usability (setup etc)."
    - Simulation Inputs and Availability of model inputs.
    - Output Usability + Model output re-usability + Conformance to open or established standards
    - To simplify the rubric, might wish to combine all the different "Reproducibility" entries into a single one. Are the model outputs likely to be reproducible to a degree sufficient for the intended user community? If no, save more output. If yes, save less output.
  - It was a challenge to ingest the list in order to determine which descriptors should be kept.

- How do we rate these? Some are very subjective and answers will depend on your mood that day vs. some that can be quantifiable.
- Who answers these questions? Different stakeholders might think differently? Does the "score" change with time?
- Get a survey from the modeling community to gain information about what they think are the minimum targets/design choices they would want
- Descriptor classes: "all other things being equal"
- Answering these questions depends on "What is the goal?" Data generation vs. knowledge generation.
- Two issues -(1) Any study with modeling data at all we need the reproducibility, (2) if you have a bigger dataset that is of value to a larger community.
- What is the message we want to send to the funders and publishers?
  - If you want to have reproducibility of the science, data is not itself sufficient. Other things needed to do and keep, rather than just a bunch of data.
- Interplay between descriptors makes it hard to talk about individual descriptors, particularly cost.
- Cameron: to solve student/advisor data management issue on publications, one option may be to identify a corresponding data manager, which may differ from the corresponding author

---

- **Compiled notes from spreadsheets about specific descriptors**
  - Computational cost
    - does this include energy cost?
    - "Cost" has disk and core hour component.
    - might be hard to quantify for an average user
  - Model/Code availability
    - Would does it mean to "be available/accessible"? Institutionally supported? Can you provide a facility to make your model reproducible?
    - Different considerations for public vs private sector.
  - Experiment Goal(s)
    - Narrow project goal(s) and additional opportunities? Sensitivity experiments?
  - (un)expected user community
    - Are usage/citation metrics going to be gathered? For an existing dataset, are there existing usage metrics? What about impacts metrics? Use in decision-making vs research?
  - Data Volume Reduction Capabilities
    - Relates to "usability" of data format/archive structure? Many dimensions -analysis ready…
  - Provenance tracking

        - challenging to develop a "general" standardized taxonomy. Maybe a subset? -workflow documentation (cylc workflow) Theoretical clean provenance vs all the actual details of workflow
    - Value to Community
        - Value is what you get from other descriptors. Maybe too vague. You will get different view points on what "value" means. How to quantify?
    - Statistical reproducibility
        - Might be hard to distinguish from feature reproducibility in some cases.
        - Combine with “feature reproducibiilty”?
    - Bitwise reproducibility
        - Difficult to achieve, though potentially possible. Not typically asked. Would need exact code, computer, hardware, compiler flags. Need to update the justification for rankings.
        - Is this really important?
    - Usability
        - "Usability" depends on the user. Might be combined with "metadata" category, and/or workflow reproducibility category.
    - How many people would be interested?
        - Hard to know until its out there. Different communities may have different expectations and needs for the data. Also related to discoverability.
    - Output availability
        - May only want a sample of the output in some cases.
    - Model metadata
        - "reproducibility" categories above are all affected by model version
    - Longevity: Usefulness
        - Take into account how data is being used. E.g. data used for decision making may want to keep longer. Could look at file access counts as an indication of usefulness.
    - Experiment reproducibility
        - How do you know there is adequate detail to do it yourself? Have to have some knowledge of process to get things to work.
    - Willingness and means to curate, maintain, and migrate as needed
        - might have willingness but not means, and for how long do you keep it? Make it “easily” accessible for 10 years but then move it to “dark” storage after that. WORS data, write once, read seldom.
    - Code usability
        - Seems like these questions are 2 very different concepts (ease of use vs private IP)
    - Incorrect (flawed) simulations
        - How does this get added later if its discovered?
    - Model type
        - more types of models. forecast, reforecast, dynamical/statistical model (LIM), analysis... it's hard to determine value based on model type

    - Updated version of experiment?
        - Consider eliminating as long as version is recorded in model metadata
    - Version control of configurations
        - May not apply to hydrology
    - Peer-reviewed results and configurations?
        - ? Isn't this what the rubric is attempting to address?
    - Model configuration / experiment/simulation setup
        - Are other aspects of setup/reproducibility subsets of this descriptor? E.g. experiment reproducibility, experiment setup, availability of documented or automated workflow, provenance tracking, model metadata, model evaluation/configuration
    - Highly Influential Scientific Assessment - See https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/OMB/inforeg/peer_review041404.pdf

---

SPECIFIC DISCUSSION

How can we facilitate different levels of participation/resources?

- How can first year grad students vs multi million dollar projects
- Motivating the use of workflows
- Reach out to your publisher if you really don't have a place to put your data
- Minimum standard for workflow and small dataset that can be supported by publisher and made accessible with the publication
- We need templates and examples to help students follow best practices more easily
- Badging system for reproducibility and ability to reproduce outputs
- Can publishers have templates for data documentation (maybe also suggest that they have a minimum bar of that is needed, for instance, a well-documented workflow, pointing to notebooks, and code, and sample datasets that can be attached to the actual publication so that they can live together rather than be separated from the publication)
- Small companies and also international players which might have different resources
- Journals can help identify resources
- There should be more impetus on the really large project to make tools that can be used by more people.
- The way to make it easier for smaller entry folks (and people who are transitioning between positions quickly- post docs) is to advise them on what standards to use as they are making their data. If people use standards they can integrated into larger repositories more easily.
- For example CMIP6 is very standards heavy -- would like to see some of this happening in the model itself
- Examples of standards - UniData
- There are a lot standards out there but people aren't implementing them uniformly
- All data CF compliant

SPECIFIC DISCUSSION

Tiers of Data:

(the objective of the “descriptor questions” is to decide how many of these things are needed, and for how long.)

(if we had infinite time, we might do data reviews more often, but it’s much better to do something like NOAA, and have a “default” discard period for each category below (with some categories, never to be discarded)

- Mandatory: Experiment Description
- Optional: Code
    - Model Base Code
    - Configured Code (validated or not) (with parameters)
        - and input data
    - (Conformance - information about how the configured code met the experiment requirements)
- Optional: Simulation Description
    - Platform Description
    - Ensemble relationship(s)
- Output Data
    - Unlikely: Tier -1: “Tuning and/or parameter setup runs”
    - Unlikely: Tier 0: Restart Dumps
    - Possible: Tier 1: Timeseries at “highest temporal resolution”
    - Possible: Tier 2: Timeseries at “appropriate reduced temporal resolution(s)”
    - Consider: Tier 2a: Subsets of Tier2 for specific periods
    - Consider: Tier 3: Statistical Summaries
        - Temporal averages
        - Reduced Spatial Resolution
    - Probable: Tier 4: Snapshots
    - Probable: Tier 5: Data to support figures
    - Probable: Tier 6: Figures

Neil: ...or…e.g., have a matrix, and classify the data/descriptor as requiring one or more. These “requirements” might be dictated by publishers, funders etc. We might agree some simulations require C3, while others might only require B0 etc. See table below for an example (what the tiers / horizons are ideally is debatable/subjective). In principle higher tiers/horizons would include the ones below. E.g. C3 would also include C0 to C2.

	Example horizons | v	Tier 0	Tier 1 data	Tier 2 data	Tier 3 data

Example tiers -->		Descriptions and configurations	Key statistical summaries	Selected timeseries	Full timeseries
Time horizon A	Until paper acceptance	A0	A1	A1	A3
Time horizon B	5 years	B0	B1	B2	B3
Time horizon C	50+ years	C0	C1	C2	C3

SPECIFIC DISCUSSION

Synopsis so far:

- There are two main groups we need to consider -
  (1) smaller or more specific studies without a broad downstream community.
    - Here there focus really needs to be on reproducibility.
    - This can be driven through requirements from journals, rather than requiring all of the data they should require that reviewers can substantiate the claims made in the manuscript and then it is up to the authors to provide a combination of workflows/data subsets in a usable way that reviewers can actually substantiate claims.
    - Generic data dumps and git hub repos with poorly documented code don't end up being used and are not actually serving this purpose.

  (2) datasets where the PI's expect a large downstream community or where they can make the case that the model results are sufficiently unique/important/novel/expensive that they need to be more fully archived.
    - To some extent the case for this needs to be made at the outset of project to the funding agency and there needs to be funding and structure in place early on to support the real costs associated with making a dataset usable.
    - If PI's make the case for data archiving they need to show they are putting resources towards it
    - Also for very large expensive datasets like this there should be some additional community outreach to decide what are the results that a broad user community might be interested in.
    - It might be the case that global and national models tend to be more likely to fall in this category because they have a broader user base.
- The issue of timing is hard - more outputs may be needed at first but down the road the core of what needs to be saved is less.
    - This is challenging because taking things out of archives is not easy

- Also this would need to happen often after projects have ended and there may not be funding.
- Making outputs more modular is important so we can pick and choose what we save for how long.

NOTES FROM PRESENTATIONS

- Doug -Restructure files and compression
    - Preserve Workflow with tool such as cylc (sp)
        - Supports reproducibility
    - Use of controlled vocabularies and standards -community agreed upon
    - ES-DOC protocol -detailed description of numerical experiment (paper has now appeared at https://doi.org/10.5194/gmd-13-2149-2020)
    - Reproduce experiment, not simulation?
    - Documentation of data and experiment is critical
    - Metadata friction -costs a lot of effort
    - Experiment documentation directions vs data dump.
    - 5 year lifetime on NCEP model output in their storage system. Not publicly available by default. Looking into providing data through a "cloud" data lake.
    - Human resource effort/costs to do any of the activities in this rubric?
- Gretchen:
    - Sheri M: CESM and CMIP
        - CMIP6- still contributing: over 20 petabytes!
        - (CMIP5 was 2 PB)
        - for postprocessing of CESM to CMIP
        - raw data not saved because of storage constraints
        - Python tool: PyReshaper (and PyConform)
            - compressed NetCDF
            - runs in *parallel
    - Bryan Lawrence
        - not all simulations should be FAIR
        - es-doc = earth system documentation
        - crucial metadata = why? what? how?
        - Reproducible
            - reproduce bitwise = who cares?
            - reproduce experiment = YES
        - metadata structure in Python
        - JASMIN supercomputer: specifically built to analyze tera/petabyte datasets
        - cost/benefit analysis
            - best practices versus practical reality
    - Jim Stagge
        - assessing reproducibility in hydrology research
        - model versioning
            - Cathy: and what if the model version has errors

        - need workflow (not just the data)
- Brian Gross
    - Director EMC (environmental modeling center)
        - transition to UFS (unified forecast system) from “ferris wheel of pain” :)
    - 5 year storage limit (~50 PB/year)
    - data lake in the cloud
        - can’t run in the cloud because restricted data can’t go in the cloud
        - security is a issue (same issue pulling in development code from github!)
- Pat Hogan
    - NCEI, NOMADS
    - before this, was at NRL and large-scale modeling group
    - focus on ocean forecasts for this talk
    - real-time and retrospective access to model output
    - ocean-nomads: ocean models (focus on currents)
    - Navy models: Fleet Numerical Meteorological Oceanographic Center = FNMOC (“fin-mock”)
    - ERDDAP: ERD data access program
        - there is a plotting component on website
    - only save analyses, not forecasts
    - 5 yrs atmos, 75 yrs ocean
- Mike Friedman
    - follow up Robert Pincus: editor for AGU “modeling journal”
        - problem is horribly heteorogeneous