Community Multiscale Air Quality Model Use Case

Compiled Aug. 2020 by Model Data RCN Virtual Workshop 2

Summary
Weighted rubric score - N/A
Category - Preserve the majority of simulation workflow outputs

- Use Case Description
    - High-level overview of the use case
        - Using the Community Multiscale Air Quality model (CMAQ, https://www.epa.gov/cmaq) to study ammonia in the atmosphere. They run WRF to then run MCIP. This data is then used in CMAQ and SMOKE. CMAQ is perturbed and run with different scenarios to find a difference. This creates lots of data in different stages of the project and results in many output files.
        - This is a NASA funded project. NASA wants others to use what is created via this project.
    - Science goals and basic workflow
        - To better estimate the ammonia emissions using NASA Cross-track Infrared Sounder (CrIS) data in conjunction with CMAQ.
        - There are multiple stages to this project, including running WRF, MCIP, SMOKE, iterating CMAQ, and processing the CRIS data.
        - The final product of this project will be a file with gridded ammonia emissions for North America, which is relatively small compared to what it took to create it.
- What use-case specific additional materials should be preserved and shared?
    - Data
        - Inputs to model
            - Description
                - None are needed to be preserved because the inputs are from NOAA and NASA data archives, and easily available.
        - Raw model output
            - Description
                - For long-term preservation - None
                    - Don't need output from intermediary stages once they are past those stages.
                - For short-term saving - They go through their version of the rubric for each model at the end of each stage. They delete some data at that point, and save some portions of the interim model output in case there may be some need to revisit it, because it would be easier than having to rerun the model. But they plan to only save it for some period of

time before deleting (with the time period depending on potential use).
        - Processed model output
            - Description
                - CMAQ emissions profile - Output data (only the ammonia-related variables, ~Gb)
    - Software
        - Model configuration
            - *Yes*
        - Preprocessing code
            - *Yes*
        - Model code
            - *Yes*
        - Postprocessing code
            - *Yes*
        - All models and scripts are Dockerized. They will probably keep these forever because it was hard to do.
    - Other
        - Documentation
            - Notes on how the model output was produced because it probably won't be possible to reproduce
            - Documentation on workflows and docker containers for each stage, e.g. notes on where input data came from, such as NOAA, and the settings/versions used for compiling the models.
        - Visualizations or images
            -
- Why should these things be preserved and shared?
    - General
        - Output data are being generated for any user
        - Takes a long time to compute and post-process the model output
        - Planning to develop an interface to allow people to select data based on geographic region, to reduce download volumes.
        - Important distinction between development runs and production runs
            - Good software engineering and documentation should enable rerunning old versions of the model if necessary.
        - May not have control over hardware, which might change and cause difficulties in recreating exact output
        - Difficulty in regenerating outputs, either by yourself or the potential users, lean toward keeping the outputs.
    - Reasons why the things listed above are important
        - Note expected/intended audience and what they expect/need
            - Are there specific people who will be using the data downstream?
                - N/A
            - Possible/aspirational users?

        - N/A
    - Note any temporal considerations, such as particular products that become more/less useful over time
        - N/A
- Could refer to individual rubric descriptors in this section - which descriptors are most important/useful to guide the preservation recommendations for each case?
    - Cost to store vs cost to rerun including labor hours