CONUS II Climate Simulation Use Case

Compiled Apr. 2022 by Model Data RCN team

Summary
Weighted rubric score - 58
Category - Preserve selected simulation workflow outputs

- Use Case Description
    - High-level overview of the use case
        - Goal is to use Weather Research and Weather (WRF) model output to better represent current and future climate simulations.
        - This is a follow-on to a previous project called CONUS-I
        - CONUS-II is a 20 year simulation at 4 km resolution over Continental United States (CONUS) and part of Southern Canada
            - The simulation does a better job at representing precip at 4km resolution. It is a good scale. Lower resolution doesn't work as well, and at higher resolutions it doesn't get much better, but the amount of data goes up significantly.
        - Two 20-year runs covering a current period from 1996-2015 and a future period from 2080-2099 over the CONUS-II domain.
        - The team is working now on the future simulations CONUS, and also working on a South American simulation of the same kind.
        - A BAMS paper on southern US describes how this model simulation represents the climate better than available observations.
    - Science goals and basic workflow
        - Create a more realistic depiction of the mesoscale terrain features, which is critical to the successful simulation of mountainous precipitation.
        - Realistically simulate the major summer precipitation producer in the central U.S. without the need of conventional cumulus parameterizations.
        - Overarching goal - to study local-scale climate change. To do actionable climate research, it is necessary to get the water cycle correct, which requires the 4km resolution.
    - NCAR data repository link:
        - https://doi.org/10.5065/49SN-8E08
- What use-case specific additional materials were preserved and shared?
    - Data
        - 274.77 TB total for one simulation (i.e., 1995-2015)
        - Inputs to model
            - For simulations of past climate - Use reanalysis for inputs - ERA5 - for initial conditions and boundary conditions.

            - For simulations of future climate - using the ensemble-mean of the CMIP5 simulations' high-end emission scenario
            - Constant WRF fields for this dataset were included in the repository: 392.46 MB
        - Raw model output
            - Not included in the repository.
            - Total data volume not preserved in a repository? (might be retained on PI's local working storage)
                - The simulations output 300+ variables
                - The simulations produced significantly higher volume than what was deposited in the repository.
        - Processed model output
            - Historical period simulation, 2D files: 119.41 TB, 1 hourly
            - Historical period simulation, 3D files: 155.36 TB, 3 hourly
                - Cleaned for redundancy
                - Separate surface variables (2D)
            - So far they have only deposited the historical years. They are still running the future years simulations.
            - WRF traditionally produces classic NetCDF. They convert these to NetCDF 4, and do level 1 compression to reduce size by 80%.
            - The RDA repository staff also do processing to generate metadata to make the data easier to find, and also enables subsetting for single variables. RDA also has a special process to do the spatial subsetting given the projection used.
                - RDA also creates DOIs for the data.
    - Software
        - Model configuration
            - The constant WRF fields for this dataset are included in the repository: 392.46 MB
        - Preprocessing code
            - WRF comes with a preprocessing system, so this was not included in the repository.
        - Model code
            - The model is WRF with some modifications.
            - Person who writes the model keeps the code. Models change over time, e.g. WRF models and libraries change. The computing environment also changes. This makes it difficult to rerun the model.
            - Re-start files allow the simulation to be restarted at any point.
        - Postprocessing code
            - Post-processing code consists of some shell scripts and NCL to manipulate netCDF.
            - These are not included in the RDA archive.

        - *COMMENTARY NOTE BY RCN TEAM: According to RCN project guidance PIs should archive and provide access to pre- and post-processing codes. We acknowledge, however, the practical and funding challenges in packaging, presenting, and archiving these codes publicly.*
    - Other
        - Documentation
            - RDA creates documentation for the data access web page and variable listings.
            - No other special documentation.
        - Visualizations or images
            - N/A
- Why were these things preserved and shared?
    - General
        - Wanted to make this data available to the university community.
        - RDA provides added value - metadata creation and DOI creation, and good services and history of providing data to the community. Other partners on the project don't have equivalent data services. RDA also provides Globus transfer and subsetting for easier data access by users. RDA staff also helps users with data access questions.
        - Differences in data archiving between CONUS I and CONUS II
            - The process was quite different. Learned many things in CONUS I
                - Created a lot of single variable files for CONUS I. Compiled data into files based on optimal data file size. Took a lot of extra work.
                - Subsequent users said they wanted all variables to run models, not single variables.
            - For CONUS II, they archived the model output with all variables in the files so users can more easily pick them up and do subsequent modeling.
    - Reasons why the things listed above are important
        - Expected/intended audience and what they expect/need
            - Are there specific people who will be using the data downstream?
                - University community
                - Partners on the project. There are 100s of partners.
            - Possible/aspirational users?
                - Users could be studying some phenomena at higher resolution, e.g. using WRF tools to use 4km resolution as boundary conditions for 1km simulations.
                - Other users are interested in single variable files, and geolocated subsetting.
                - RDA enables any other users to access the data.
        - Note any temporal considerations, such as whether particular products become more/less useful over time

            - N/A
- Broader Impacts:
    - How will output from this project be used by stakeholders?
        - Had never been data at 4km resolution for South America before. Global models may not be resolving weather phenomena that happen at that scale.
    - How were stakeholders involved in the data curation decision-making?
        - Affinity group created to share workload of Cheyenne computing hours dedicated to this simulation.
    - How will stakeholders be compensated for their participation in the data curation decision-making process?
        - N/A
- Do you have any concerns about misuse of your data or software? If so, what concerns do you have, and what are the reasons for those concerns?
    - Not discussed