NARCCAP

An Assessment of the NARCCAP Data Archive

Seth McGinnis and Linda O. Mearns

National Center for Atmospheric Research, Boulder, CO

website: http://www.narccap.ucar.edu

ABSTRACT

The North American Regional Climate Change Assessment Program, NARCCAP, is a high-resolution regional modeling program for the study of climate change in North America. It involves driving different regional models with boundary conditions from different global models, all using the same resolution, spatial domain, and temporal coverage. In addition to enabling the exploration and quantification of model uncertainty, output from the models is intended to be used by the community in service of three distinct purposes: further downscaling to even higher resolution; intermodel comparison and analysis of model performance; and impacts analysis, including use in decision-making.

This poster provides an overview of the NARCCAP data archive, and discusses notable features of the archive and their impacts, both positive and negative, on the goals of the program. Such features include: format specification, standards compliance, data organization, variable selection, prioritization of results, primary vs derived data products, publication strategy, and content from multiple sources.

OVERVIEW

NARCCAP is not yet complete, but so far we have published 17+ terabytes of data in more than 22000 files. 26 out of 30 primary simulations have been run, and more than 500 registered users have downloaded tens if not hundreds of terabytes of data for use in dozens of papers.

Because it is easier to identify flaws that you encountered than bullets that you dodged, this poster will focus largely on things we wish we had done differently, but overall, the NARCCAP data archive has been quite successful in distributing data to end users in support of the program's research goals. If you need to share data from a large modeling project, the NARCCAP model is one pattern you can follow.

GOALS

- Investigate and quantify uncertainty in regional model projections.
- Develop high resolution regional climate scenarios for impacts analysis.
- Evaluate regional climate model performance over North America.

Target Uses

Impacts Analysis

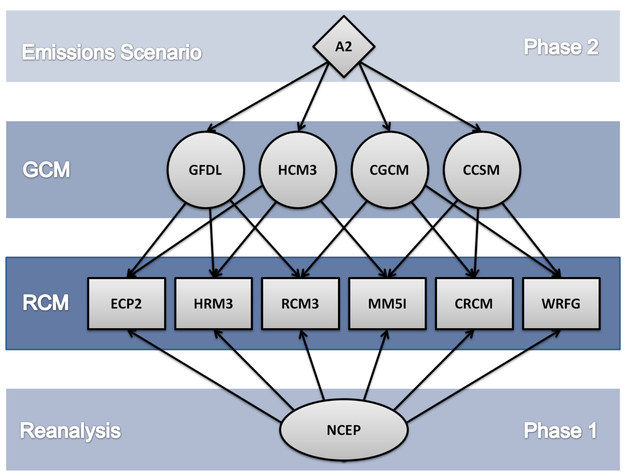
Model Evaluation

Further Downscaling

EXPERIMENTAL DESIGN

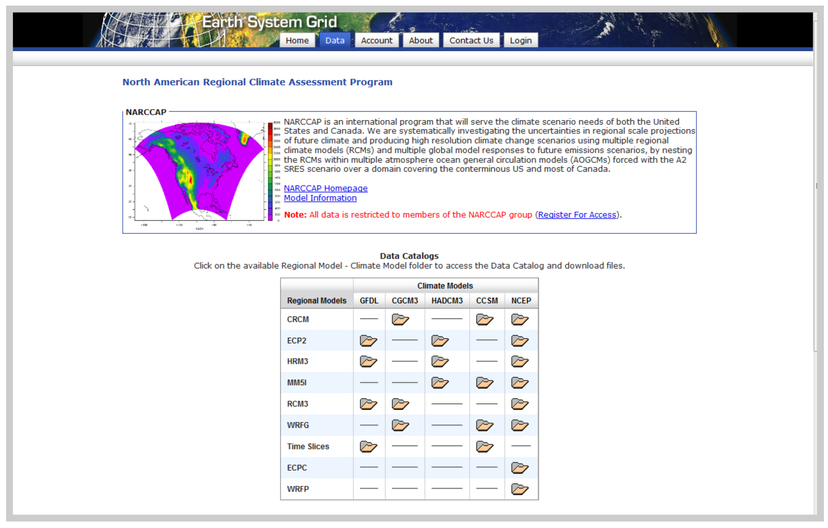
6 RCMs nested in 4 GCMs, plus NCEP-driven runs

Fractional factorial design due to funding constraints



ARCHIVE ORGANIZATION

NARCCAP data is published through the Earth System Grid (ESG) data portal. At the top level, files are organized by RCM and driver (GCM or NCEP)



Screenshot of top-level ESG page for NARCCAP data

Within each of the 30 runs, the variables are grouped together into tables by structure:

Table 1:	2D	daily	4 vars
Table 2:	2D	3-hourly	7 vars*
Table 3:	2D	3-hourly	24 vars*
Table 4:	2D	static	6 vars
Table 5:	3D	3-hourly	7 vars**

*Table 2 has the 2D variables needed for impacts analysis, while Table 3 has additional 2D variables needed for model assessment, largely following CMIP recommendations.

**Table 5 data is further divided up by pressure level.

This organizational scheme mostly aids data production, but has some utility for end-users.

It might have been useful to further subdivide Table 3 thematically (e.g., creating a separate table for radiation variables) to streamline post-processing and make it easier for users to find data of interest.

Static data should have been Table 0; because it's the same from run to run, it doesn't really fit in, but sits off to one side from the rest of the data.

PUBLICATION

NARCCAP data checking and publication is centralized. Modelers at different institutions post-process the model output and send it to NCAR, where it is quality-controlled and published. Centralization is slower than distributed publishing, but results in more consistent data quality and was necessary for publication via the ESG data portal.

We chose to prioritize the publication of data based on its usefulness to the end-users. Because impacts users are the most populous sub-community, we focused on publishing impacts-relevant variables (Table 2) first. The table organization of variables helped with this.

Because there was so much of it, we also split the 3D data in Table 5 into two groupings by pressure level (primary and secondary), to get the data of greater interest out the door as quickly as possible.

Some data, like the static Table 4 variables, didn't fit well into the organizational scheme of the data portal, and was published on the NARCCAP website instead.

All registered users are subscribed to a mailing list that is used for official NARCCAP communication. Whenever a new dataset is published, we announce it to the mailing list. Data is typically published in large batches, so mailing list traffic is infrequent.

ERROR NOTIFICATION

Occasionally, we discover problems with published data. Because access to NARCCAP data is restricted to registered users, ESG tracks file downloads. Thus, we can contact every affected user directly and inform them when replacement data is available. This is a dramatic improvement over an errata page, which may not be seen by those who already downloaded the erroneous file, while at the same time being needlessly alarming to those who come later.

One flaw with the ESG publishing system is that it does not distinguish between major and minor revisions to published files (i.e., changes to the primary variable data vs changes to metadata and ancillary variables). This sometimes requires special action to handle, causing delays in the publication of corrected data.

STANDARDS COMPLIANCE

NARCCAP data is stored in NetCDF format conforming to the CF metadata standard. In addition, it adheres to a custom project spec that follows the CMIP specifications.

Standardization is a major boost to usability across the board. The more uniform the data is in structure, the less time a user has to devote to understanding how to interact with it. Standards also enable the use of smart tools that can automatically handle low-level details. For example, NetCDF data files that meet the CF requirements can be read into the latest version of ArcGIS with no extra work required to convert formats. (So long as the CF standard is followed stringently.)

On the data production end, following a standard eliminates a lot of decision-making about data representation, and allows the use of third-party format checking tools. Requiring the model output to be even more uniform than we did would have saved work in the long run, and could have been accomplished by an up-front investment in the creation of template files to be used in post-processing.

SPLITTING FILES

To stay under the 2 GB file size limit imposed by older versions of NetCDF, we split the data from each run into 5-year chunks. This caused or exacerbated a great many problems, and in hindsight it would have been much better to have only one file per variable per run, and instead to push users to update their netcdf installations.

We originally planned to publish Table 5 data as 4D blocks, but changed course and regularized the file structure by splitting the data up into separate files for each pressure level. This simplified post-processing and QC, and enabled us to prioritize the publication of important levels. So splitting files isn't always bad.

VARIABLE CHOICES

We prioritized impacts-relevant variables for greater usability. However, it would also have improved usability to store the most commonly-used variables (temperature and precipitation) using the most commonly-used units (°C and mm/day) rather than canonical units (°K and kg/m^2/s).

Snow-water-equivalent (swe) is more useful than snow depth (snd). Inadvertently, NARCCAP requirements include the latter but not the former.

Table 5 includes two cloud-composition variables. These are useful for model analysis, but perhaps not enough to warrant the storage space they require. To save space, we did leave pressure height (zg) off Table 5, because it can be computed using the hypsometric equation. However, this proves to be a false economy; zg is needed for most 3D analysis, and calculating it is time-consuming and inexact.