Lessons and Pitfalls in Archiving Large Datasets: The NARCCAP Experience

Seth A. McGinnis, Larry R. McDaniel, and Linda O. Mearns

ISSE, National Center for Atmospheric Research, Boulder, CO

email: mcginnis@ucar.edu *website: http://www.narccap.ucar.edu*

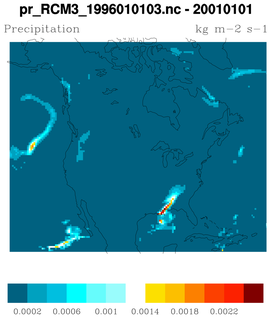
ABSTRACT

The North American Regional Climate Change Assessment Program (NARCCAP) is an international program to produce high resolution climate change scenarios and investigate uncertainties in regional scale projections of future climate by nesting multiple regional climate models (RCMs) within multiple atmosphere-ocean general circulation models (AOGCMs) forced with the A2 SRES scenario and with historical data over a domain covering the conterminous United States and most of Canada and Northern Mexico.

The resulting datasets will total roughly 60 terabytes in size and must be archived for distributed storage and made available to global change impacts researchers worldwide via the Earth System Grid (ESG). This presentation will describe our data management procedures and the lessons we have learned about handling such a large flux of data, maintaining its quality and integrity, and ensuring that the final product is usable by the impacts community, GIS practitioners, climate analysts, modelers, policy-makers, and other end users. The importance of data formats, metadata standards, and flexible tools for visualization, checking, and automation will be discussed, as well as social and other significant factors.

AUTOMATED ANALYSIS

This plot of the last timestep in a datafile was generated by automated QC tools. To aid in finding problems, visualization uses raster plotting instead of smooth contours, contour intervals based on min and max values, and annotation information pulled from the file metadata.



THE ARCHIVING PIPELINE

After the climate model has been run, there are still multiple steps involved in transforming simulation output into usable data.

Modeling: NARCCAP simulations are run by modeling groups of 1 to 3 investigators at various institutions. Each model is run by a different group.

Post-Processing: Simulation output is converted to CF-compliant NetCDF according to the NARCCAP output spec document. Format-checking the output for correctness was a slow and iterative process on the first, NCEP-driven runs, but will be faster for the GCM-driven runs.

Data Transfer: The original data management plan calls for data to be loaded onto 1-TB hard drives and shipped to LLNL, but alternate methods have been used in the cases where they are available and faster.

Backup: After the data has arrived at LLNL and before it is checked or changed in any way, it is backed up to the NERSC HPSS for disaster recovery purposes. Keeping the original copy of data untouched is a good general practice for error recovery.

Quality Check: The NARCCAP QC team performs basic checks of data integrity, checking that the files are correctly formatted and contain data that appears to be valid. Minor metadata errors (which are common) can be corrected without requiring the modelers to reprocess output. Commonly-requested "value-added" derived data products, such as total precipitation aggregated at the monthly scale, are generated as part of the QC process.

Archive: A copy of the data is made to the NCAR Mass Store for archival purposes.

Publication: Data ready for release is copied to a suitable server and published to the Earth System Grid data portal website. A new version of the ESG portal is in development that will provide many user-requested features, including spatial and temporal aggregation and subsetting, improved search, and on-line visualization capabilities.

Correction: Some data will inevitably be later found to have problems, whether by modelers or by end users. Because registration is required to access NARCCAP data, users who have downloaded the affected files can be contacted individually and informed of the problem and the availability of corrected data.

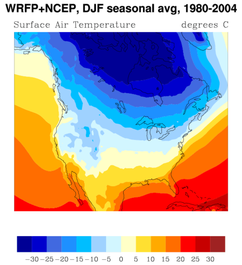
NARCCAP OVERVIEW

- 4 different AOGCMs driving 6 different RCMs
- 50 km spatial resolution
- 3 hourly temporal resolution
- 52 output variables
- 2 high-resolution GCM timeslice experiments
- Future scenario: A2 SRES emissions

RCM \ GCM	Phase I	Phase II			
	NCEP	GFDL*	CGCM3	HADCM3	CCSM*
CRCM	DONE		1		2
ECPC	DONE	1		2	
HRM3	DONE	2		1	
MM5I	DONE			2	1
RCM3	DONE	1	2		
WRFP	DONE		2		1

Example Data: Seasonal Climatology

WRFP+NCEP, DJF seasonal avg, 1980-2004



Phase I: RCMs are driven by historical (1979-2003) observed (NCEP2 Reanalysis) data

Phase II: Each RCM is driven by 2 GCMs for current (1968-2000) and future (2038-2070) scenarios. GCM/RCM pairings are chosen for maximum value in statistical analysis.

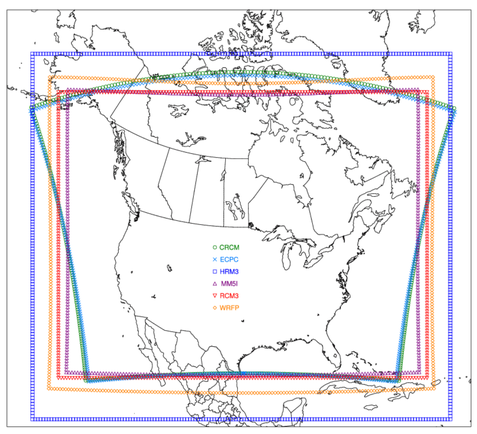
Timeslices: Atmospheric components of the GFDL & CCSM global models are run at 50 km resolution using observed SST data (offset in the future scenario) instead of a coupled ocean.

CURRENT STATUS: Impacts-relevant data for Phase I has been published to ESG for all RCMs. All groups have finished Phase I modeling, and 3 of the 6 groups have finished their first GCM-driven run. Both timeslice experiments have been run, and all data from the GFDL timeslice has been published via ESG.

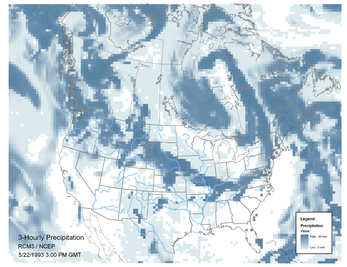
INTECOMPARISON, MAP PROJECTIONS, & SPONGE ZONES

Although all six RCMs have the same 50 km spatial resolution and cover the same domain, because they use different map projections they cannot cover exactly the same set of gridpoints. The models also differ in the size of the 'sponge zone' where the model data is mixed with the driving lateral boundary conditions. This results in differences in the effective domain size, and in one case required adjustment of the simulated domain to increase the area of overlap. In addition, certain of the required output variables simply do not exist for some models, or cannot be captured as single variable. These factors all show that even in an experiment designed to enable the intercomparison of different models, there are limits on how much they can be made to model "the same thing". This issue has spurred the NARCCAP team to invest in the development of tools for interpolating the data to different grids.

NARCCAP RCM Domains



GIS, METADATA, & IMPACTS USABILITY



GIS Compatibility
This map, created in GIS, shows precipitation for a 3-hour period on the morning of May 22nd, 1993. On this day, heavy storms caused severe flooding in Sioux Falls, South Dakota.

Members of the impacts community use GIS heavily in their analyses, so GIS compatibility is very important for making the data useful to them. ESRI's ArcMap v.9.3 supports direct import of NetCDF data if the files follow the CF metadata standard. To achieve GIS interoperability, the NARCCAP team has had to ensure that file metadata stringently follows the CF standard, particularly with regard to specification of the map projection parameters used by each model. This effort was significant, but has paid off in transparent ingestion of NARCCAP data into GIS for the impacts community, which illustrates the importance and value of using a standard and adhering to it strictly. CF-compliant NetCDF data can be interpreted by a variety of other programming and analysis tools as well, including NCL, R, IDL, and Matlab, and can be exported to plain-text files readable by spreadsheet programs.

DEVELOPING AND SUPPORTING A DIVERSE USER BASE

The user base for high-resolution climate change scenario data spans a broad spectrum of technical sophistication, from climate researchers, who require information about the details of model configuration, at the high end, to members of the general public, who need data summaries presented in a form suitable for consideration as one factor among many in general policy decision-making.

Impacts users occupy a middle ground. Providing them with effective support requires making relevant data easily accessible. One approach that greatly furthers that goal is to prioritize data processing according to the utility of the end product, so that the most used data becomes available soonest. This has the added benefit of getting data out and into the hands of users, who will perform a more thorough testing and inspection of the data than the modelers and publishers can, before the entire data stream has been processed, allowing corrections to be folded into the archiving process.

Promoting usability may necessitate the development of ancillary and derived products, such as precipitation in cumulative form rather than as a rate. However, it is difficult to anticipate the needs of a large and diverse user base, and therefore often good policy to hold off on creating such products until it is clear there is a real demand for them. Much valuable information can be collected by cultivating vanguard users, who are granted access to data in the early stages of availability.

TECHNIQUES FOR MANAGING LARGE DISTRIBUTED PROJECTS

The biggest challenge faced by a project involving collaboration between multiple institutions is its disconnected nature. Geographical and organizational separation of the collaborators induces a kind of friction that slows interactions and introduces error. Frequent, clear, and effective communication is the best tool for counteracting this effect, but a distributed project will always proceed at a slower pace than a consolidated one. Specific techniques of value include:

Restrict decision-making to involve the minimal set of participants needed to effectively address the issue.

Automate everything you possibly can, including communications with users.

Develop procedures for dealing with the normal flow of data and information.

Don't hesitate to use *ad hoc* substitutions for procedure when they are more effective.

Combat email fatigue by deploying wikis and other technologies that match medium to message well.

Employ higher-bandwidth communication channels when dealing with more complex issues.

Recognize that the entire process is a balancing act, and be prepared to make changes when things go wrong, as they inevitably will.