Appendix N: Climate Scenario generated for the DOE Office of Science High Performance Network Planning Workshop

The past

The last 10 years have seen a significant increase in the complexity of the climate models used to study natural and anthropogenic climate change. In the early 1990s, a typical climate model was an early attempt at an integrated system to simulate the then-suspected effects of human activities on Earth's climate, using relatively coarse resolution (in space and time) components representing only the two or three most readily simulated physical parts of the climate system – the atmosphere, ocean and simple sea ice. The atmospheric component of this period had a horizontal resolution of approximately 800 km by 500 km, and only 9 vertical levels. Such models typically generated a few gigabytes (10^9) of data from a small number of short (less than 100 years model time) experiments. The data volume from a single model year was on the order of one to three gigabytes. The data were stored at the site where the supercomputing capability resided, on low capacity magnetic tapes. The data were not often transferred to climate researchers located away from the supercomputing center, and Internet security protocols were minimal or nearly non-existent.

Today

Current state-of-the-art climate models, such as the PCM (Parallel Climate Model) and the CCSM (Community Climate System Model) utilize considerably higher resolution simulations of more components (atmosphere, ocean, sophisticated sea ice, and land surface) of the climate system. The atmospheric model, for example, has 300 km by 300 km horizontal resolution, and 18 vertical levels. A given model is typically integrated many dozens of times (up to 1000 years of model time for a single model run) to enable climate scientists to better statistically analyse the model results. This process can easily generate tens of terabytes (10^{12}) of data, both from the model output and from the derived results. Both types of data are frequently transferred from site to site, where climate scientists and their support staff engage in their analysis.

This process of data transfer often strains the current capacity and robustness of computer networks. Over the last three years, experience with the PCM has shown that it is possible for the data transfer to take longer than the generation of the data, which can critically affect subsequent data analysis and visualization. Specifically, since 1998, the PCM has been used to create approximately 15 terabytes of model output data, from nearly 10,000 years of model time. 10 terabytes of that total has been transferred to the PCM's primary data archival center, which is PCMDI at Lawrence Livermore National laboratory. Projects such as the Earth System Grid are engaged in R&D aimed at increasing the efficiency and effectiveness of the data transfer and retrieval process. The approach is essentially twofold: work faster using traditional approaches and work smarter using new ones. In the first case, we want to increase the efficiency and speed of large-scale data transfers using new Grid technology with provisions for TCP/IP tuning and parallel streams. In the second case, we focus upon developing abstracted object-level distributed data services that facilitate the movement of optimally small portions of remote data. These developments will be complemented by effective metadata and replica approaches, server-side data reductions and data analyses, and intrinsic secure transmission properties the help address the much more stringent security protocols now in place at many computing centers.

The next five years

Over the next five years, climate models will see an even greater increase in complexity than that seen in the last ten years. Influences on climate will no longer be approximated by essentially fixed quantities, but will become interactive components in and of themselves. The North American Carbon Project (NACP), which endeavors to fully simulate the carbon cycle, is an example. Increases in resolution, both spatially and temporally, are in the plans for the next two to three years. The atmospheric component of the coupled system will have a horizontal resolution of approximately 150 km and 30 levels (see figure 1.) A plan is being finalized for model simulations that will create about 30 terabytes of data in the next 18 months, which is double the rate of model data generation of the PCM.

These much finer resolution models, as well as the distributed nature of computing resources, will demand much greater bandwidth and robustness from computer networks than is presently available. These studies will be driven by the need to determine future climate at both local and regional scales as well as changes in climate extremes - droughts, floods, severe storm events, and other phenomena. Climate models will also incorporate the vastly increased volume of observational data now available (and that available in the future), both for hindcasting and intercomparison purposes. The end result is that instead of tens of terabytes of data per model instantiation, hundreds of terabytes to a few petabytes (10^{15}) of data will be stored at multiple computing sites, to be analyzed by climate scientists worldwide. The Earth System Grid and its descendents will be fully utilized to disseminate model data and for scientific analysis.

2007 and beyond

In the following five years, climate models will again increase in resolution, and many more fully interactive components will be integrated. At this time, the atmospheric component may become nearly mesoscale (commonly used for weather forecasting) in resolution, 30 km by 30 km, with 60 vertical levels. Climate models will be used to drive regional scale climate and weather models, which require resolutions in the tens to hundreds of meters range, instead of the typical hundreds of kilometers resolution of the CCSM and PCM. There will be a true carbon cycle component, models of biological processes will be used, for example, simulations of marine biochemistry (which affects the interchange of greenhouse gases like methane and carbon dioxide with the atmosphere), and fully dynamic vegetation. These scenarios will include human population change and growth (which effect land usage and rainfall patterns) and econometric models, to simulate the potential changes in natural resource usage and efficiency. Additionally, models representing solar processes, to better simulate the incoming solar radiation, will be integrated. Climate models at this level of sophistication will likely be run at more than one computing center in distributed fashion, which will demand extremely high speed and tremendously robust computer networks to interconnect them. Data volumes could reach several petabytes, which is a conservative estimate.

Conclusions

In summary, climate change science will create far larger data volumes than is currently common, and the rate of data volume creation per calendar year will continue to increase. The history of the growth of data stored on the NCAR mass storage system (figure 2) illustrates this issue.This explosive growth demands greatly increased network speed, bandwidth and robustness.

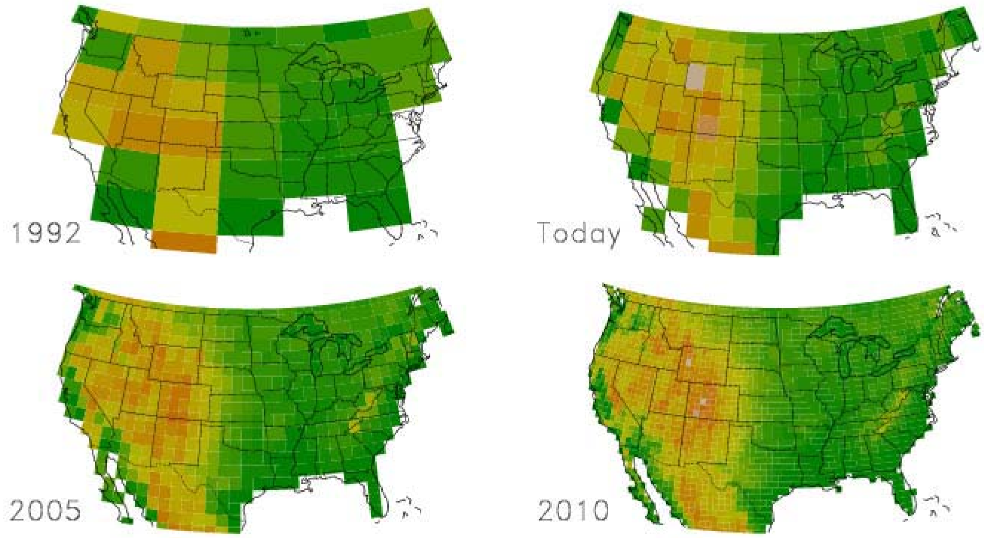

Figure 1. The increasing resolution of climate models

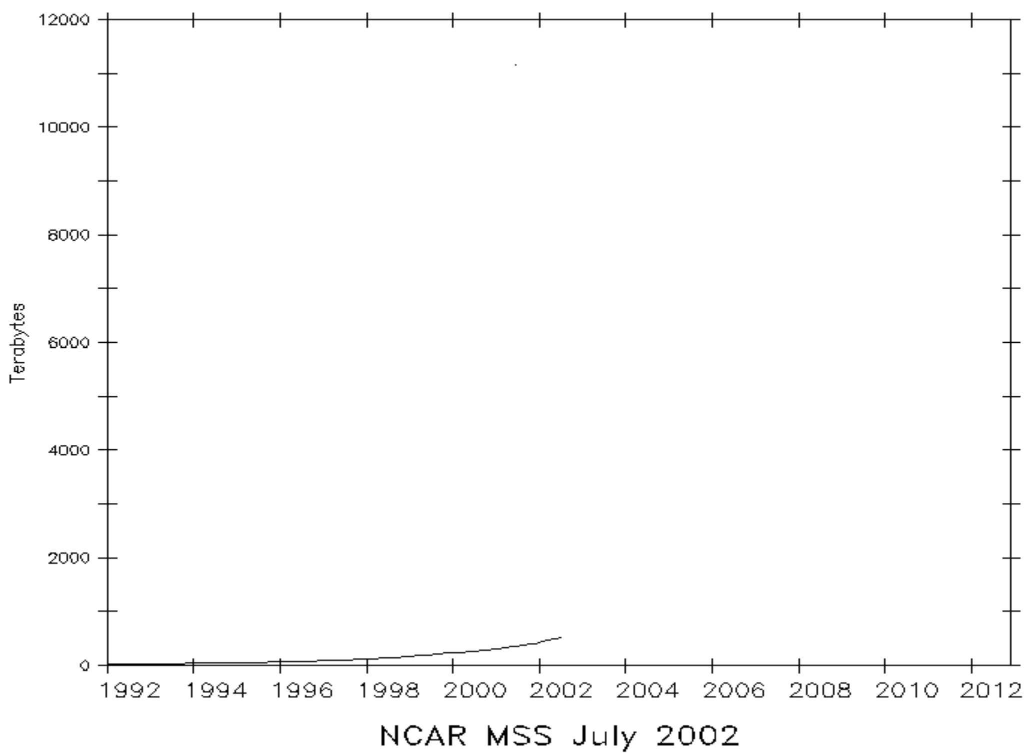

Figure 2. NCAR mass storage growth, 1992 to 2002 (courtesy Gene Harano, NCAR)