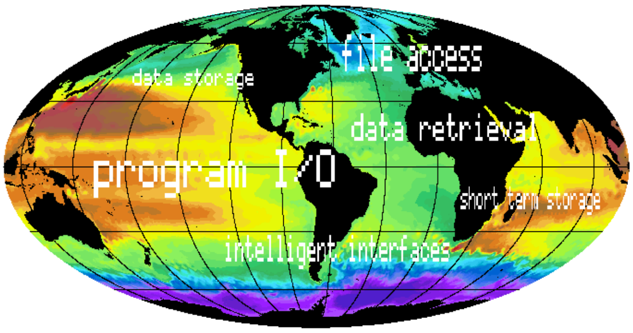
Data Management in Climate Research

by
Kerstin Kleese, CLRC - Daresbury Laboratory
Email: k.kleese@dl.ac.uk, URL: http://www.dci.clrc.ac.uk/Group/DCICSEHPA



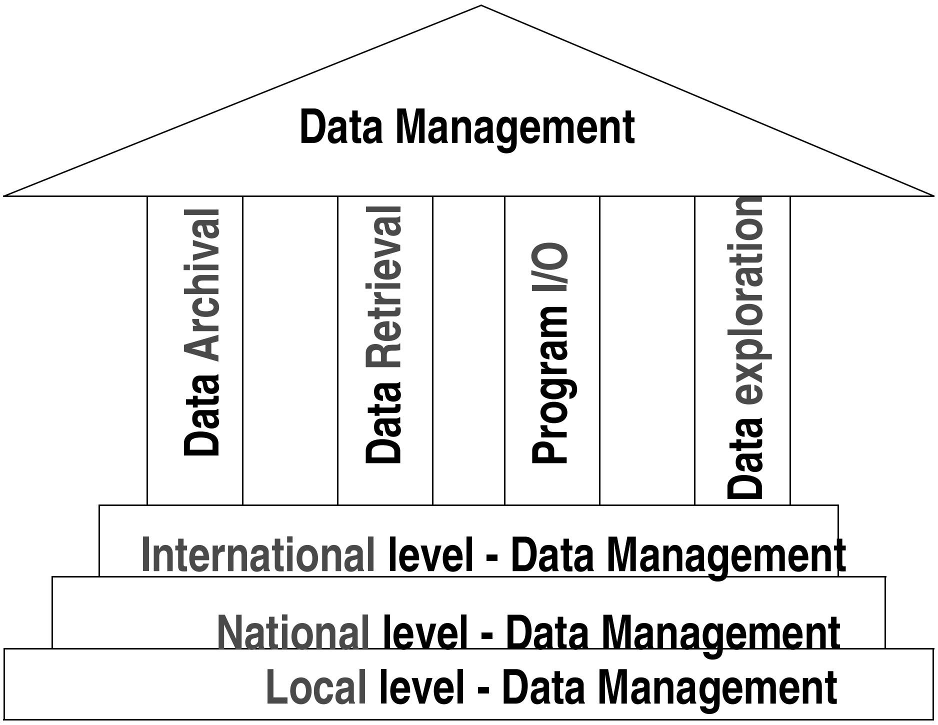
To address the identified environmental research challenges, researchers need access to a wide range of observational data and model output, covering the human, physio-chemical and biological components of the Earth system. This data should be recognized as a highly valuable resource, but data from existing data centres are frequently under-used because of the difficulties of accessing the data and of assessing whether they contain anything of interest or relevance [Environmental Change Network, 1997]. Furthermore data exploration is still in its early stages.

The complexity of the phenomena and processes involved in climate research, make powerful super-computing essential. Today's parallel models are some of the most demanding codes we have. They push the machines available to their limit and are still in need of more resources. However the bottleneck in running these codes is not so much the code performance as the data handling strategies employed.

The seriousness of the problem is increasing rapidly, as new computer and observation technologies encourage the production of more data in shorter time spans. For an efficient working environment, mechanisms have to be put in place to support scientific work more effectively. In response, CLRC – Daresbury Laboratory has set up a new project to investigate this important issue and to develop new more satisfying strategies [Kleese K, 1998]

What is Data Management?

Data Management can be described by the software and hardware techniques used to facilitate the management and the exploration of data.


Data Management
Data Archival
Data Retrieval
Program I/O
Data exploration
International level - Data Management
National level - Data Management
Local level - Data Management

We can than group these techniques into four different categories:

Data archival mechanisms
Data retrieval mechanisms
Support for program I/O
Data exploration techniques

Further consideration has to be given to local, national and international aspects of these points as well as the question of whether we are dealing with "approved data sets" or "private data". Approved data sets are validated data sets stored for public access, e.g. in a national data centre.

LOCAL LEVEL

Data archival and retrieval techniques comprise managing the stored data, mechanisms for access and additional user interfaces (e.g. graphical interfaces/ web access)both for private data and approved data sets. Support for program I/O includes "in-time" delivery of input data at program start and during the program run, mechanisms to handle program output during and after the program run, and programming techniques to generate suitable I/O patterns in the code. These points are highly dependent on computer system type and configuration. Data visualisation is a special subsection of program I/O. It includes "in-time" and "real-time"" delivery mechanisms for input data for the visualisation tool as well as mechanisms to handle data output during and after visualisation. Data formats also have to be considered, which ones are most commonly used, and whether they are available or easily compatible etc. Data exploration refers to tools that help the presentation of information in an accessible way as well as techniques that generate new information through analysing the existing data, e.g. statistics, pattern recognition or improve the data quality e.g. data fusion.

NATIONAL AND INTERNATIONAL LEVEL

On the national or international level we are more interested in organisational aspects like the similarity or rather dissimilarity of:

- Data formats
- Access mechanisms
- Computing environments
- Languages

We are also looking for mechanisms to manage and organize our personal data, as well as a more general support for locating data sets.

IN GENERAL

To determine the quality of the Data Management at a particular site or level, you do not only have to look at the quality of the involved components, but more importantly at how smoothly they work together and how user friendly the whole solution is. Only this will determine how effectively and efficiently it can be used.

Data - A valuable resource

Progress in climate research is based on two components the scientists and the scientific data holdings. Data are the lifeblood of scientific research and represent therefore one of our most valuable assets. In their day to day work researchers need efficient access to a wide range of observational data sample collections and model output covering the human physio-chemical and biological components of the earth system.

> ***"Up to date, data from existing data centres are frequently under-used because of the difficulties of accessing the data and of assessing whether they contain anything of interest or relevance."***
>
> **[Environmental Change Network, 1997].**

Data Centres

The technologies we use in Climate research have undergone a considerable change new observation, computation, and visualisation systems have been developed over the years, and more and more data is produced faster than ever before. This is putting high demands on the data centres! They have to prepare the data for further use by the academic and industrial community, often under considerable time constraint. And if they are linked to a computing centre they also have to manage the requests for private data archival and retrieval.

"National data centres, like world data centres, typically have a discipline focus, thus, which data centre you approach depends on the type of data you’re interested in."

Ann Linn, ICSU Panel on World Data Centres.

Another aspect that we need to consider is, that traditionaly many data centres focus on a specific scientific area, e.g. sun spots. As a result todays climate data are distributed over numerous systems and sites. The UK alone has at least 37 different data centres. 37 different data centres also mean 38 different access mechanisms and even more data formats in which the data are stored. What it doesn’t mean is, that there is any support to find these data centres, assess the stored data and explore it. And this is only on a national level! Many of nowadays research projects need access to international data holdings and they are also combining a growing number of scientific disciplines. Existing arrangements are unsuitable to support these types of projects efficiently.

Climate Modelling

Computational environmental modelling has steadily progressed over the last decades. Modern super computing technology enables us to produce increasingly realistic representations of environmental processes. Research groups all over the world have developed sophisticated parallel codes, which explore the capabilities of current High Performance Computing platforms very well. So, everything should be allright. Unfortunately that is not the case. As funny as it may sound, but the speed and the efficiency of these models is also their downfall. Todays climate modelling codes produce significant amounts of output and they require in many cases large input data sets.

The following tables originate from a paper by Alan O’Neill and Lois Steenman-Clark, UGAMP [O ’Neill A, Steenman-Clark L, 1998]:

Table 1. Typical atmospheric experiment data sizes for 10 model year run with data output four times per model day.

Spatial Resolution	Data Sizes (Gbytes)
climate	36.5
seasonal	74.8
climatology	98.1
forecast	324.1

Table 2. Typical ocean experiment data sizes for a 10 model year run with data output once per model day

Spatial Resolution	Data Sizes (Gbytes)
4° x 4° (global)	5.8
1° x 1° (Atlantic)	21.5
1° x 1° (global)	133.2
1/4° x 1/4° (Atlantic)	347.1

So, how do todays systems cope with these requirements? The capabilities of components like: processors, networks, compilers, and scientific libraries have improved dramaticly over the last years. But disk I/O and data archival and retrieval mechanisms haven't kept pace with that. In fact, they allready represent a major bottleneck for codes with high I/O requirements.

"The data volumes generated by climate models can be very large and are a problem to deal with especially when the models generating this data have been optimised to run quickly and efficiently on HPC platforms"

Alan O'Neill and Lois Steenman-Clark, UGAMP [O'Neill A, Steenman-Clark L, 1998].

"A major bottleneck facing the SEA model now, is that of I/O."

Matthew I. Beare, University of East Anglia [Beare MI, 1998].

Future Tendencies

The seriousness of the situation is increasing rapidly, as new computing and observation technologies encourage the production of more data in shorter time spans.

And lets be honest who isn't interested in increased model resolutions, longer runs, more complex models and larger ensembles. So, the volume of the data we are going to produce in the future and therefore the amount of data the data and computing centres have to cater for, will increase dramaticly over the next years.

For an efficient scientific working environment better mechanisms have to be put in place to support scientific work on a local, national and international level!

"By the end of the century the volumes of data available will increase by several orders of magnitude."

Peter Churchill, CEO [Churchill P, 1995]

"The European Centre for Medium Range Weather Forecast (ECMWF) expects that its already big data store doubles in size every 18 months!"

Dick Dixon, ECMWF [Dixon D, 1998]

Scope of the Project

In the short term we will continue to study and assess the current situation of Data Management in Climate Research, as well as its influence on everydays scientific work. A big part of our efforts over the next year will be directed at an detailed investigation of all major hardware and software components involved in Data Management. We want to assess their quality, ease of use and how well they work together with other products. We would also like to get a better overview about Data formats. How many are there? Which ones are most commonly used? What is influencing peoples decision to use a particular format?

In the long run we would like to build on the the results of the previously described work. The study of the current situation will give us indications on, which areas of Data Management have the biggest influence on the scientific work and should therefore be the first ones to be targeted. The survey of the hardware and software components, will enable us to give advice, on preferable product choices. Often the exchange of a single product or the addition of a new one can result in remarkable improvements. We also hope to establish an information service that keeps the community up to date on the afore mentioned topics as well as new developments.

However our biggest goal will be, to be able to investigate and develop new tools:

To provide better access to the data, through standardized access interfaces and better support for multidisciplinary searches.

But more importantly we would like to develop tools to further the easier, better ,and more efficient exploration of data. We think that this is an area that could provide tremendous benefits for the community.

Some projects, though only covering limited areas, already show their potential, if they could be applied to a wider area.
There is for example Data fusion, where data from different sources and with different properties, e.g. Data from the Hubble space telescope and from a ground based telescope, are automaticaly combined on the computer to get the best of both worlds. This technique applied to other data could provide us with generally higher quality input data.
Another idea is the application of data mining techniques on large amounts of stored data, in this case to investigate air polution tendencies, without the necessity of model runs. It is thereby freeing valuable computing time.

Summay

Data Management influences many areas of our daily working live, wether we want to store data, retrieve data, if we are concerned with program I/O optimization or if we want to explore the existing data holdings, again and again we are confronted with Data Management mechanisms.

And as Data Management has such a big influence, we should all be concerned about its quality. Today all of us have to deal with more and more tasks that have nothing to do with our real scientific work. In many cases we have to accept this development. But many of the tasks related to Data Management are just unneccessary! There are better ways, there are Easier ways and there are faster ways to handle them!

So lets make sure that this will change in the future.

References

Beare MI (1998) The Southampton - East Anglia (SEA) Model: A General Purpose Parallel Ocean Circulation Model. In: Allan RJ, Guest MF, Simpson AD, Henty DS, Nicole DA (ed) High Performance Computing. Plenum Publishing Company Ltd., London, p 339-348Churchill P (1995) Centre for Earth Observation (CEO). In: The Globe on the Net, Issue 25

Dixon D (1998) ECMWF weathers enormous growth with help from IBM. ADSTAR. In: Distributed Storage Manager customer experience, WWW

Environmental Change Network(ECN) (1997) Baseline data and information from long-term monitoring sites in the UK. WWW

Kleese K (1998) Data Management in Climate Research. In: ERCIM News, Issue 34Konstandinidis, Georgios and John Hennessy (1995) MARS - Meteorological Archival and Retrieval System User Guide. In: ECMWF Computer Bulletin B6.7/2, Reading

Linn A (1997) The World Data Centre System. http://www.ngdc.noaa.gov/wdc/wdcmain.html#wdcMenochet, Annabelle(1998) How to use the BADC. http://www.badc.rl.ac.uk/how/

O'Neill A, Steenman-Clark L (1998) Modelling Climate Variability on HPC Platforms. In: Allan RJ, Guest MF, Simpson AD, Henty DS, Nicole DA (ed) High Performance Computing. Plenum Publishing Company Ltd., London, p 317-326