briefing paper

digital preservation europe

World Data Center for Climate: Preservation of Earth System Model Data

Increasing computing capabilities for the production of Earth system model data has created new challenges for its long-term archiving and preservation. Not all the data produced on the system can be stored in the long-term archive. A new archive concept developed by the ICSU (International Council for Science) World Data Center for Climate (WDCC) separates data storage, including an expiration date at the scientific project level, and the documented long-term archive. This new system produces a completely documented, quality checked long-term archive with a searchable data catalogue.

Introduction

The size of the data output from earth system models depends on spatial resolution, temporal output intervals, the number of variables and the data storage format. Technological developments have brought about increased computational power leading to finer spatial and temporal model resolution and the integration of additional physical and chemical processes into models. Although the costs for storage per TB decrease continuously, this is not keeping pace with increases in computational power and connected data production rates. In addition to the financial implications of data preservation, the cost of quality assurance and securing usability increases with the total size of the mass storage archive. Therefore it is not feasible to store all data produced in the long-term data archive and the long-term archiving strategy has to be modified to limit the expansion of the long-term data archive.

Strategy for Long-term Archiving and Interdisciplinary Data Utilisation

This strategy separates data into two categories or life cycle stages. The first category or stage includes project data management at the level of scientific projects with a limited lifetime, the second category or stage includes data suitable for long-term archiving. This separation ensures an aware, scientific decision to move data from the first category to the second for long-term archiving, taking into account and balancing both the rigours of good scientific practise and the limited availability of resources.

As the data transmitted to the WDCC are typically the final results from scientific projects on which scientific publications are based, the rules of good scientific practice require that this data must be available and accessible for at least ten years in order to allow for future verification of the published results. This data is also part of the general knowledge creation process and may be used in interdisciplinary scientific fields. This dual application aspect of long-term archiving, especially interdisciplinary re-use, requires even higher standards of data preservation, quality assurance and usability than data at the scientific project level. In addition to this, as final results and data products cannot be easily reproduced after a number of years and data users from a wider interdisciplinary audience may not have the necessary skills to deal with numerical model data, there is a requirement on the long-term archive to store documentation metadata along side the scientific data.

Expected life cycles of these two standard data categories are reflected in their data expiration dates. Data reaching its expiration date will be removed from the system after a warning is issued. Data identified as requiring further archiving will be moved to the long-term archive level providing a preservation lifecycle of ten years and longer according to library rules. Data which does not require further storage is removed from the archive in coordination with the data owner.

Data moved to the long-term archive level will be stored on the tapes of the DKRZ (Deutsches Klimarechenzentrum) mass storage system together with complete documentation and a second copy for security. The WDCC with its searchable data catalogue is currently building the core of this archive level. For the next generation of computer servers the WDCC documentation concept will be expanded to cover all data in this hierarchy level either in files or in the database system, allowing the long-term data archive to be completely documented and searchable.

Further information

The WDCC provides data and services for Earth system research with specific emphasis on numerical modelling and related observations. The WDCC, founded in 2003, is operated by the Model and Data Group at the Max-Planck-Institute for Meteorology in cooperation with the German Climate Computing Centre in Hamburg.
Presently the WDCC archives and disseminates more than 340 Terabytes (TB) of earth system model data and related observations. All WDCC data are accessible by a standard web-interface (http://cera.wdc-climate.de). The WDCC has approximately 1000 named users and in 2007 experienced more than 650,000 data downloads totalling 200 TB. The WDCC also works towards international cooperation in developments of networking and archive federation.

The primary data publication service is offered by the WDCC as an additional service for data of general interest which should be referenced directly in scientific publications and which are then searchable in standard library catalogues together with scientific publications. The primary data publication process has been developed together with the Technical Information Library, Hannover (TIB), and has been implemented in the STD-DOI profile (Scientific and Technical Data - Digital Object Identifier, http://www.std-doi.de).
Primary data which are identified as independent data entities in the context of scientific literature are suitable for the STD-DOI publication process. These data, together with their metadata, pass a review process and a process of quality assurance before the metadata for electronic publication and a persistent identifier (DOI/URN) are assigned. The metadata for electronic publication and the corresponding identifier are registered in the library catalogue of the TIB. Now the published primary data can be searched and accessed together with standard scientific publications. The metadata of the STD-DOI profile are open for harvesting and for integration in alternative information systems.

Data Preservation, Quality Assurance and Enabling Usability

The new and expanded long-term archiving concept of WDCC and DKRZ is based upon the need to provide a more comprehensive stewardship of data in the long-term data archive level. This naturally limits the amount of data which can be handled.

Bit stream preservation will be secured by additional tape copies at a separate location and in a different format. In addition to this a maximum number of tape accesses is set and the total number of tape accesses is recorded. If the maximum number of accesses is reached the data from the original tape is migrated to a new tape and the old one is removed from the tape silo.

Providing quality assurance for numerical model results is rendered complex by of the vast amount of data ingested and stored. It is not feasible to inspect every single number in the output. To address this challenge spot tests are performed at various levels of complexity. Quality assurance of numerical model outputs requires semantic and syntactic examinations. The semantic examination entails examination of the behaviour of a numerical model itself compared to observations and other models. This forms the major part of the scientific evaluation process. The syntactic examination considers the formal aspects of data archiving and ensures that the archiving is as free of errors as possible. This includes examining the consistency between the metadata and climate data, the completeness of the climate data and related metadata, the standard range of values, and the spatial and temporal ordering.
Although in the majority of cases they can be performed automatically, these proofs are time consuming. Nevertheless they are necessary to ensure confidence in the data archive.
Quality assurance in the WDCC and the DKRZ long-term archive is performed as a three stage process:
• the semantic check is performed at the scientific level during the runtime of the corresponding scientific project in order to decide on the validity and usefulness of the actual model results. A positive decision is the basic criterion to migrate the data from the project data management into the long-term data archive;

• data documentation and syntactic proof routines are conducted during the data integration process into the long-term archive;
• the most complete data examination is accomplished in connection with the STD-DOI data publication process. The publication process includes more detailed review and quality assurance of metadata and climate data.

The usability of the long-term archive will be improved by a complete searchable documentation of the climate data entities in the catalogue system of the WDCC. Additionally the WDCC offers web-based data access for those climate data which are stored as individual two dimensional fields in the database tables. Presently the WDCC offers web-access to more than 340 TB of climate data which are stored in 120,000 database tables and in six billion individual table entries. The average size of a single table entry is calculated to 60 KB and this corresponds to the level of granularity of data access offered.

The usability of data must also be supported on the technical level. Archive technology transfer must be downward compatible to keep old data technically readable in future. The related software should also be migrated to new platforms. Data processing tools and data format access libraries are continuously needing to work even with older data from the long-term archive.

Conclusions

The WDCC concept for preservation of Earth system model data presented here improves the long-term archive reliability within existing funding limitations.
Due to the exponential growth rate in data production, it was determined to focus resources on data stewardship, such as archive documentation, bit-stream preservation, quality assurance and enabling usability, instead of on the technical handling of vast amounts of data.

References

[1] ICSU World Data Center Climate (WDCC): http://www.wdc-climate.de

[2] German Climate Computing Centre (DKRZ): http://www.dkrz.de

[3] Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I., Wächter, J.
Data Publication in the open access initiative Data Science Journal, Vol. 5, p79-83, 2006.

[4] Lautenschlager, M., Stahl, W.
Long-Term Archiving of Climate Model Data at WDC Climate and DKRZ
In: E Mikusch (Ed.): PV2007 - Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data, Conference Proceedings. DLR, German Remote Sensing Data Center, Oberpfaffenhofen, 2007

[5] Toussaint, F., Lautenschlager, M., Luthardt , H., World Data Center for Climate Data - Support for the CEOP Project in Terms of Model Output
Journal of the Meteorological Society of Japan, Vol. 85A, pp. 475-485, 2007

Michael Lautenschlager, *Model and Data / Max-Planck-Institute for Meteorology, Hamburg, Germany*
michael.lautenschlager@zmaw.de