World Data Center for Climate Data—Support for the CEOP Project in Terms of Model Output

Article *in* Journal of the Meteorological Society of Japan Ser II · February 2007

DOI: 10.2151/jmsj.85A.475

Frank Toussaint
Deutsches Klimarechenzentrum

World Data Center for Climate Data—Support for the CEOP Project in Terms of Model Output

Frank TOUSSAINT, Michael LAUTENSCHLAGER, and Hans LUTHARDT

Max-Planck-Institut für Meteorologie/World Data Center for Climate, Hamburg, Germany

(Manuscript received 14 February, in final form 13 November 2006)

Abstract

The World Data Center for Climate (WDC-Climate) is hosted at the Max Planck Institute for Meteorology (MPI-M) in Hamburg, Germany. WDC-Climate stores global and regional model output for the CEOP project, and raw model output is available in a number of different data structures. This model output is currently being restructured into a more homogenous form. This paper provides an overview of the CEOP model output and its overall data structure at WDC-Climate, as well as the different ways that this output can be accessed.

1. Introduction

CEOP (Coordinated Enhanced Observing Period) was launched to investigate fundamental aspects of the Earth's global water and energy cycle. The full CEOP data system merges data from different sources as completely as presently possible for CEOP Phase 1 (October 2002 to December 2004). The CEOP data include earth-observation satellite data provided by Space Agencies, model output from weather-prediction models run by national weather-prediction (NWP) centres, and locally observed station data from 35 CEOP reference sites (in-situ data). These three different data classes are managed by the CEOP data centres: satellite data are archived at the University of Tokyo, Japan (http://ceop.tkl.iis.u-tokyo.ac.jp/data/download.html), in situ data is archived at the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory (EOL) in Boulder, USA (http://data.eol.ucar.edu/codiac/), and global and regional model output are archived at the World Data Center for Climate (WDC-Climate) run by the Model and Data Group (M&D) at the Max Planck Institute for Meteorology (MPI-M) in Hamburg, Germany (http://mad.zmaw.de). A single user-interface for these data classes is provided by the Committee on Earth Observation Satellites (CEOS) and the Working Group on Information Systems and Services (WGISS) Test Facility for CEOP (WTF-CEOP) based at the Remote Sensing Technology Center (RESTEC) in Tokyo (Fig. 1).

CEOP data management emphasises interoperability between the contributing data archives and the different data classes. The overall goal is to provide for more effective comparisons between in situ observations, satellite data, and data from weather prediction models, as well as between the models themselves. CEOP's pilot application is the global water and energy cycle, but the CEOP data archives could potentially be used for many other process-based studies in the field of climate research. Once the CEOP Phase 1 database has been completed, it will provide a unique dataset for the global water and energy cycle over a designated period of time.

Examples of scientific applications and descriptions of additional infrastructure aspects

Corresponding author: Frank Toussaint, Max-Planck-Institut für Meteorologie/World Data Center for Climate, Model&Data Group, Bundesstrasse 53, Hamburg, D-20146 Hamburg, Germany.
Email: Toussaint@dkrz.de

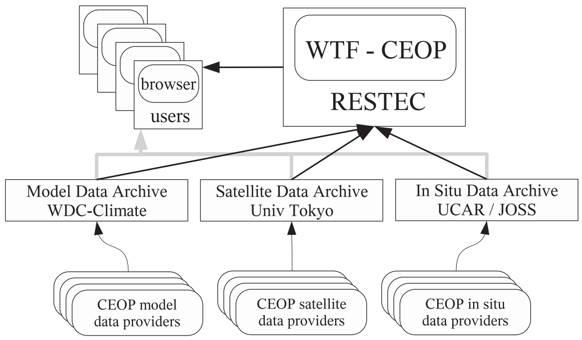

Fig. 1. The three CEOP data centres store satellite, model, and in situ data; the WTF-CEOP provides an integrated view of all data sources. Users are able to obtain direct access to the data at the data centres.

are provided by other contributions to this special CEOP issue of the Journal of the Meteorological Society of Japan. For more detailed descriptions of CEOP data management, see the data-management website at http://www.eol.ucar.edu/projects/ceop/dm/. In the present paper, we describe the WDC-Climate's archive of model output.

2. World Data Center for Climate

2.1 *World Data Centers worldwide*

Founded in 1931, the International Council for Science (ICSU) is one of the oldest non-governmental organizations in the world. On the occasion of the International Geophysical Year (IGY) in 1957/58, ICSU established the World Data Center (WDC) System, which today includes 52 centres in 12 countries. These data centres provide scientific data for various topics in earth and space sciences. Detailed information on the WDC System, including a list of centers, can be found at one of the WDC home pages, e.g., http://www.wdc.rl.ac.uk/wdcmain/.

In 2003, the German WDC Cluster for Earth System Research was established by three German WDCs: WDC for Marine Environmental Sciences (WDC-MARE, Bremen), WDC for Remote Sensing (WDC-RSat, Oberpfaffenhofen), and WDC-Climate (Hamburg). Collaboration is ongoing with a proposed WDC of the Lithosphere (WDC-TERRA, Potsdam). All four data centres operate long-term data archives for Earth System Research and related projects such as CEOP.

2.2 *Mission and structure of WDC-Climate*

The WDC-Climate was approved by ICSU in 2003 and is run by M&D of the MPI-M. M&D is funded by the German Government and makes use of the computing and storage facilities of the German Climate Computing Centre (Deutsches Klimarechenzentrum, DKRZ).

The mission of WDC-Climate corresponds with the basic tasks of M&D. Data for Earth System Research are collected, stored, and disseminated to serve the scientific community. Emphasis is placed on climate modelling and related data products. Close cooperation with data centres in corresponding fields such as earth observation, meteorology, oceanography, palaeoclimate, and environment either exists or is planned in order to eventually establish a complete network for earth system data.

A brief description of WDC-Climate can be found on the organization's home page (http://www.wdc-climate.de/) or in the Guide to the World Data Center System (http://www.ngdc.noaa.gov/wdc/europe/climate.html).

3. Data collection for the CEOP project

3.1 *Contributing data centres*

Twelve institutes that cover nearly all of the continents contribute output from Numerical Weather Prediction (NWP) models to the CEOP model output database. Most of the model output covers the time from the first annual cycle of the CEOP to the end of 2004; however, output from the National Centers for Environmental Prediction (NCEP) is currently being provided for the period from 2005 to the present day.

Table 1 provides an overview of the promised CEOP model output; however, not all of the promised output has been delivered to WDC-Climate and not all of the delivered output has been entered into the CERA database. Figure 2 provides the present status of data assimilation; a current version of this graph is available on the World Wide Web (http://www.mad.zmaw.de/projects-at-md/ceop/).

3.2 *Gridded output*

For the CEOP project, it was agreed to use WMO-GRIB format for all model output on spatial grids; however, this does not imply that all

Table 1.

Institute	NWP Models
Bureau of Meteorology (BoM) Research Centre (BMRC)	Grid M3
Centro de Previsão de Tempo e Estudos Climáticos (CPTEC)	CPTEC/COLA
European Centre for Medium-Range Weather Forecasts (ECMWF)	Operations, ERA-40 (and continuation)
Experimental Climate Prediction Center (ECPC)	Reanalysis-II, Seasonal Forecast Model, Regional Spectral Model
Epson Meteo Centre (EMC)	EMC model
NASA Global Land Data Assimilation System (GLDAS)	Mosaic Land Surface Model, Noah Land Surface Model
NASA Global Modeling and Assimilation Office (GMAO; formerly DAO)	GEOS3, GEOS3-TRMM, GEOS4
Japan Meteorological Agency (JMA)	JMA-GSM for operational global data assimilation system (3D-Var)—three different resolutions
Meteorological Service of Canada (MSC)	GEM (Global Environmental Multiscale model)
NOAA National Centers for Environmental Prediction (NCEP)	Global Reanalysis I, Global Forecast System
National Centre for Medium Range Weather Forecasting (NCMRWF)	NCMRWF analysis-forecast system
United Kingdom Met Office (UK MetOffice)	Global Unified Model

model output are homogeneous. Output is provided by different centres using different data structures. Differences exist in terms of the vertical scale; parameter lists ("code tables"), time scales of forecast and assimilation, and data-storage structures. Further information on the nature of these differences is provided on the CEOP website (http://www.eol.ucar.edu/projects/ceop/dm/model/model.html), which provides links to specific NWP documentation.

At WDC-Climate, model output is mostly stored as time series of individual variables at a specific altitude level. This system of data storage is used because it satisfies the majority of the requirements of Earth System Research. The CEOP model output provided by the NWP centres, however, contains all parameters and levels at one point in time. The first step is to store these as-is data directly into the database and then make them publicly available as four-dimensional raw data. In a subsequent step, it is planned to convert the four-dimensional data to a two-dimensional single parameter time-series structure. This will significantly improve the access performance and will allow for better integration of the gridded output into the integrated WTF-CEOP (see Fig. 1).

3.3 Model Output Location Time-Series

The gridded model output is stored as time series of two-dimensional (latitude, longitude) global data at increments of 3, 6, 12, or 24 hours. To compare the model output of each of the NWP centres with in situ measurements, Model Output Location Time Series (MOLTS) at the specified locations of reference site were extracted from the global grid model output by the NWPs. The resulting one-dimensional curves refer to the 41 originally defined CEOP MOLTS locations that correspond in turn to 35 nearby measurement stations that contribute to the CEOP observation data (in situ data; see http://data.eol.ucar.edu/codiac/; a list of the station characteristics for the reference sites can

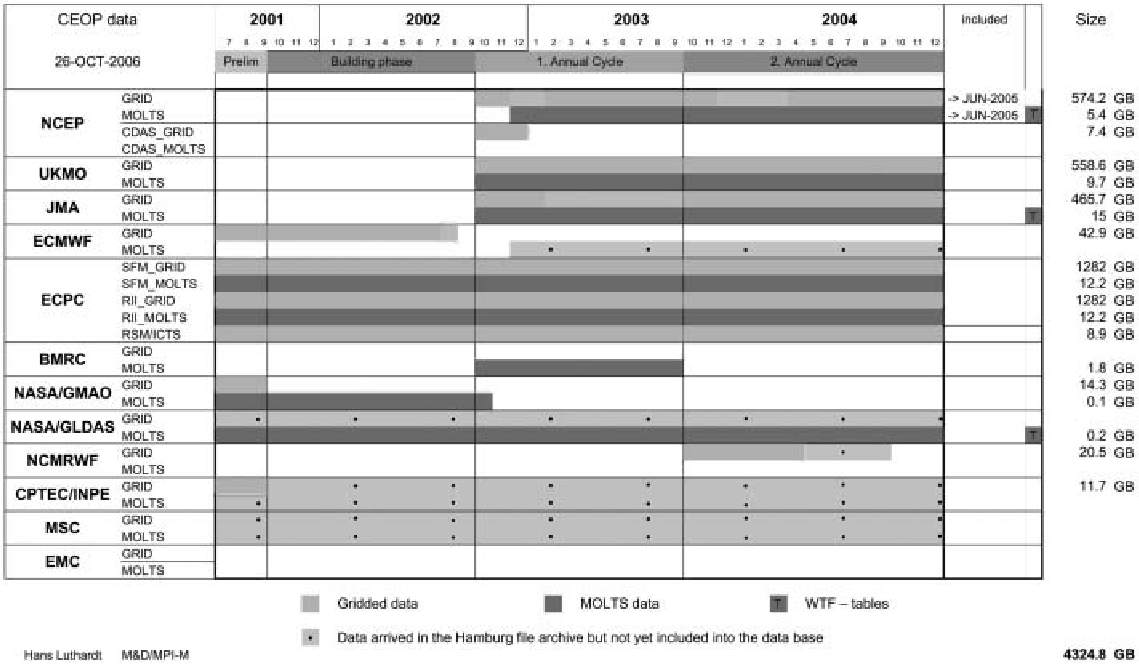

Fig. 2. The status of the assimilation of CEOP data into the database at WDC-Climate, as arranged in terms of prediction centre and data type (grid/MOLTS).

be found at http://www.eol.ucar.edu/projects/ceop/dm/documents/rsite/, and the locations are summarized at http://www.eol.ucar.edu/projects/ceop/dm/documents/ref_site.html).

For MOLTS, it was initially decided to house the model output in the following three formats. First, the original data blocks are stored in the format in which they are provided to WDC-Climate by the CEOP partners. These original data are disseminated in different formats. Depending on their source, they vary from pure binary to different versions of the Network Common Data Form (NetCDF) and plain ASCII. In addition, the descriptions of the contents (metadata) are not always comparable. Most data in this raw format are now accessible to the CEOP community via the WDC-Climate web browser.

Second, a copy of the MOLTS is stored in a homogenized form as NetCDF-CF (NetCDF Climate and Forecast Metadata Convention; http://www.cgd.ucar.edu/cms/eaton/cf-metadata/) standard format, i.e., these data are stored in NetCDF format, and every file contains the CF-required keywords in the file header. Support from the CEOP Community, especially from the Bureau of Meteorology Research Centre (BMRC, Australia) and GKSS-Forschungszentrum (Germany), was required to perform these conversions. It should be noted that these reformatted MOLTS are not yet included in the WDC-Climate database in NetCDF-CF format because the conversion process is incomplete.

Finally, a third copy of the MOLTS is directly entered into database tables as standard real numbers, from which it is accessible number-by-number via the Internet. This form of storage is optimized for access from the WTF-CEOP system (WGISS Test Facility for CEOP, Distributed Data Integration System Prototype) and from other CEOP partners who wish to realize automated online data access. The interface, which is necessary for downloading table data from the database, is now running, and data retrieval from WTF-CEOP began in November 2005 for test data.

3.4 Data structures and homogenization

To receive, archive, present, and disseminate data is not the only task of a data centre. Users commonly need to retrieve and download the data in a structure and size that are different from those of the uploaded data. WDC-Climate usually receives grid data as large files in which all levels and parameters of a single time step for the entire globe are stored in a single data file. Most NWP centres store their model output in this form. Without subsequent data processing, this NWP form of data storage results in large individual table entries (up to 20 megabytes).

Most users ask for a customized data structure for their data download. In particular, a time series of globally gridded values of one parameter at one altitude level is usually requested. Data tables then contain individual time series with small (tens of kilobytes) table entries. For performance reasons, it is not possible to retrieve these time series on the fly from the original model datasets because the data quantities are too large to store permanently on disks: additional data processing is required.

In most cases, providing access to only the coarse data blocks of the raw data forces the user to download and handle much more data than is required for the application. This mismatch between raw data storage and data application dictates that the data should be reconfigured according to the likely application prior to storing the data in the database system.

In general, the CEOP grid model output spans the following dimensions: the source model, the physical parameter, space, model-initialization (assimilation) time, and forecast step. The coverage of this seven-dimensional space differs among different models. Figure 3 provides an example of four of the data structures, showing their different distributions in assimilation time and forecast step.

The current state of the ongoing assimilation process of CEOP model output (grid and MOLTS stored in the original data structures) into the WDC-Climate database can be found at the WDC-Climate web site, where a graphical timeline indicates the temporal coverage of available data (http://ceop.wdc-climate.de).

4. Data storage and access: CEOP model-output database

4.1 Data storage as binary objects

Within the CERA-2 (Climate and Environmental Data Retrieval and Archiving) system, data are stored as binary large objects (BLOBs) in database tables. BLOBs are the smallest entities that can be handled by the database system. This usually means that each time step is stored as a single table entry in the Gridded Binary format used by the World Meteorological Organization (WMO-GRIB); however, all other data-storage formats can also be archived, with NetCDF currently emerging as the preferred format. The concept of BLOB-storage provides efficient access to application-adapted data entities that vary in size from several kilobytes to several megabytes. In CERA-2, the data entities usually contain one time step of gridded output of one or more physical parameters.

All BLOBs are administrated by the CERA-2 database, which provides the necessary metadata, including access information, and allows for detailed data-access control.

4.2 Graphical user interface of the CERA-2 system

WDC-Climate offers online catalogue access to the metadata via a Graphical User Interface (GUI, Fig. 4) that can be accessed using any web browser. At the server, this is achieved using Java Server Pages (JSP) and servlets, i.e., no additional plug-ins are needed on the client side and problems with firewalls are avoided. This user interface enables anonymous data searches on the basis of physical parameter, scientific project, and various other topics that are defined in the metadata. To access and download the actual data requires a WDC-Climate personal user account that can be requested by e-mail to data@dkrz.de.

In the archive catalogue, the data are grouped into projects, one of which is CEOP. The data in each project consist of various groups, which in the database structure are called *experiments*. For the CEOP project, the global model output from one NWP is separated into at least two experiments: MOLTS and gridded output.

Each of these experiments is then subdivided into datasets that represent the next-lowest level in the data hierarchy. For example, in the case of CEOP MOLTS, the datasets represent the output at 41 MOLTS locations.

For two-dimensional (grid) time series, data processing can be performed before the data are downloaded. By selecting the regional area,

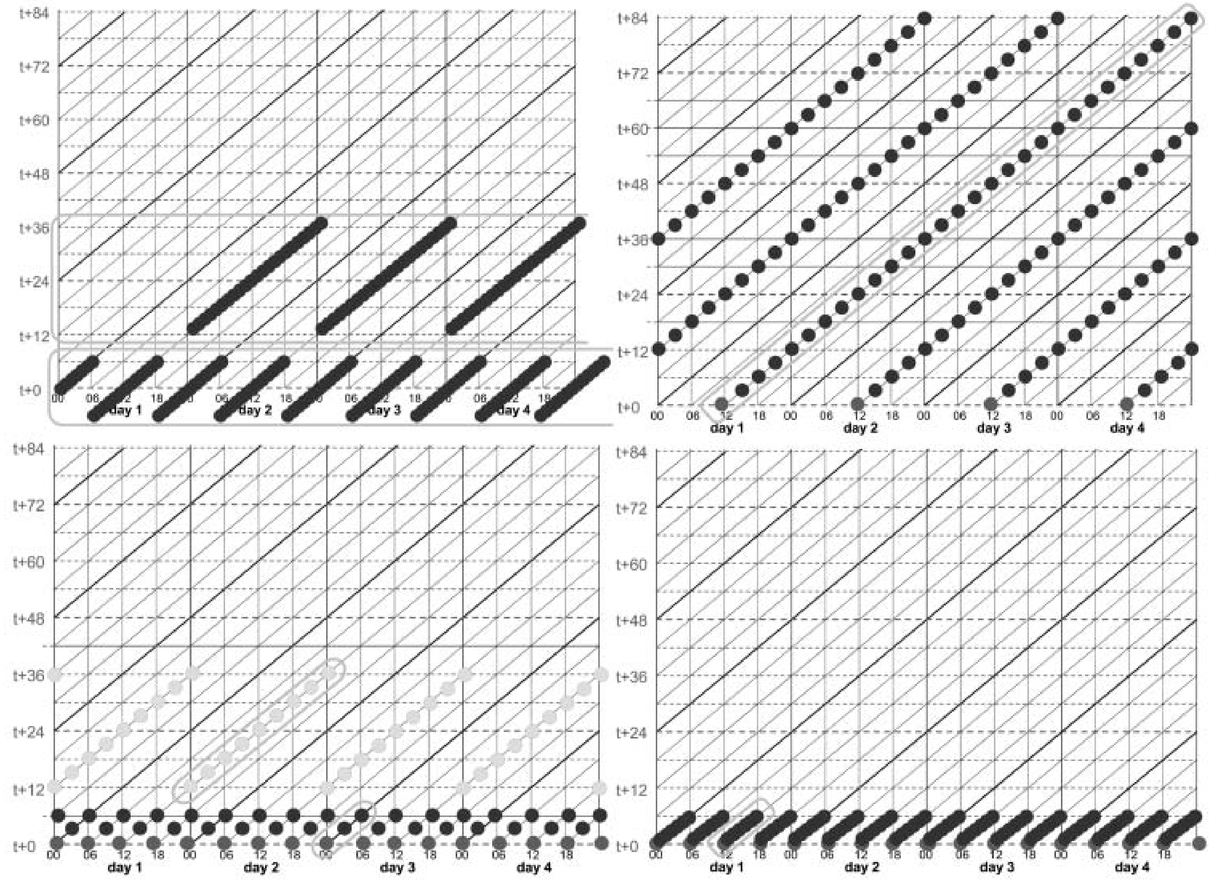

Fig. 3. Examples of different data structures for CEOP data. The graphs refer to data from NCEP (upper left), JMA (upper right), BMRC (lower left), and UKMO (lower right). Diagonal lines connect points of equal model runs/initialisations (equal assimilation time), while horizontal lines connect points of equal forecast steps, and vertical lines connect points of equal verification time (= assimilation time + forecast step). Dots represent available data of one or more physical parameters at one or all altitude levels per dataset. Dots that correspond to a single raw data file are placed within a green surround. For a complete list of CEOP model raw-data structures, see the section titled "Structure of the model output" in the MPI-M CEOP website at http://www.mad.zmaw.de/projects-at-md/ceop/.

the user is able to reduce the data download volume for all data in WMO-GRIB format. Format conversion from WMO-GRIB to compressed text format can also be performed.

4.3 Selection of CEOP model output from the WDC-Climate catalogue

Direct access to CEOP model output is provided by the CERA-2 metadata catalogue at WDC-Climate (http://cera.wdc-climate.de/), where the user can browse (Fig. 5) the available data. The selection of a CEOP project yields a scrollable list of approximately 23 different CEOP experiments, including the original data files as supplied by the NWP centres (raw data).

The selection of any of the CEOP experiments provides detailed information on the data by clicking on the "experiment information" button. The other button leads to the subordinated datasets for the experiment. To download data to the user's local disk, the dataset must be added to the process list, which serves as a shopping cart.

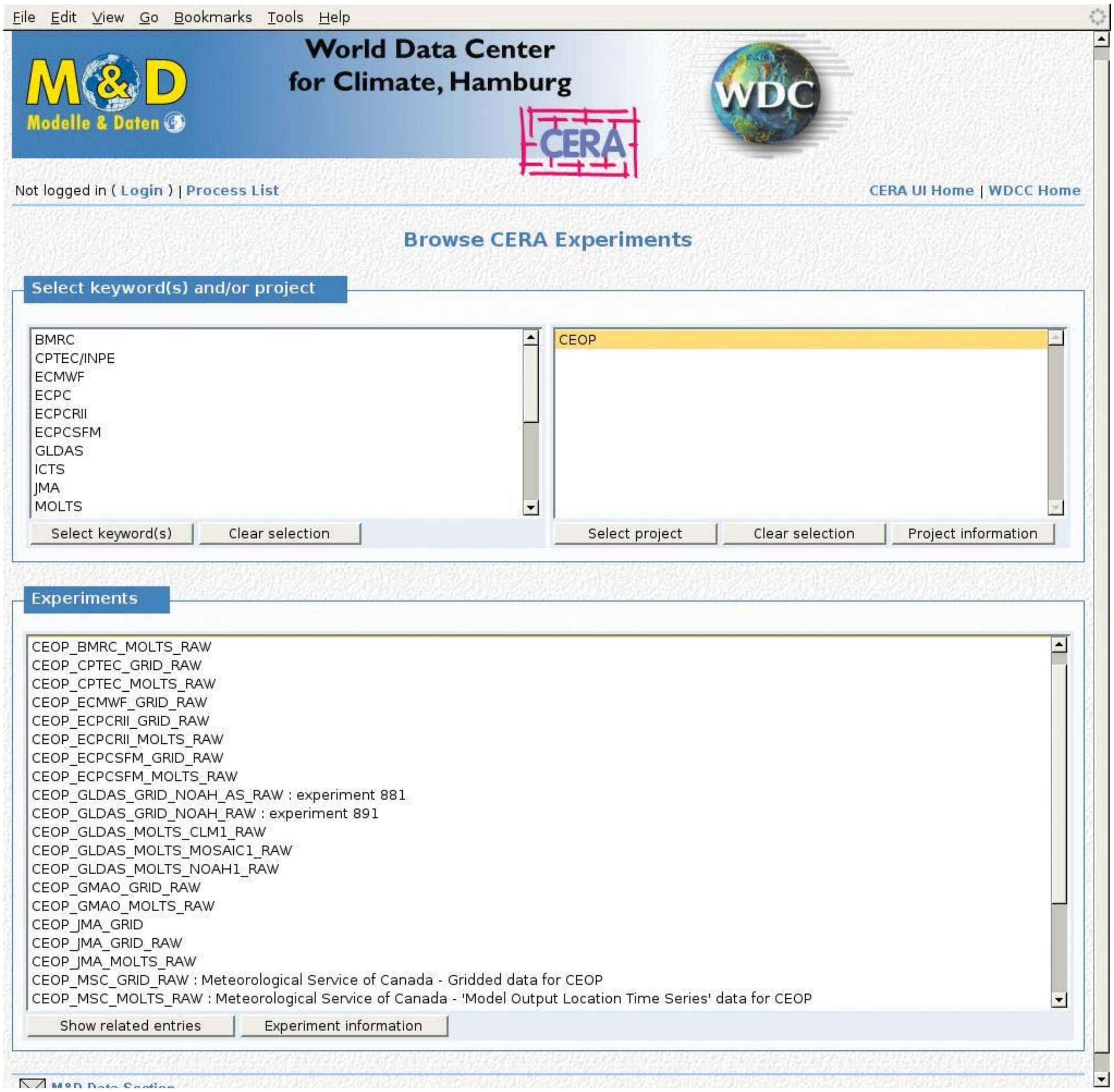

Fig. 4. The Graphical User Interface of WDC-Climate can be accessed by any web browser (http://cera.wdc-climate.de). On the upper right it provides project selection (e.g., "CEOP"), which leads to the experiment list below. The experiments' names consist of the computing centre's acronym and the data type (e.g., grid/MOLTS). All names ending with "_RAW" indicate data that was originally transferred to WDC-Climate; other entries refer to homogenized data. If more than one data stream is provided, a specification is appended to the centre's acronym (see Table 1, Section 3.1 for a list of different data streams).

The floppy-disk icon in the process list leads to the web page where the region and the start and end time of the requested data can be specified for download. From here, the data stream is initiated from the WDC-Climate disk cache to the local disk; however, for some purposes a download via script may be more appropriate (see Section 5.5).

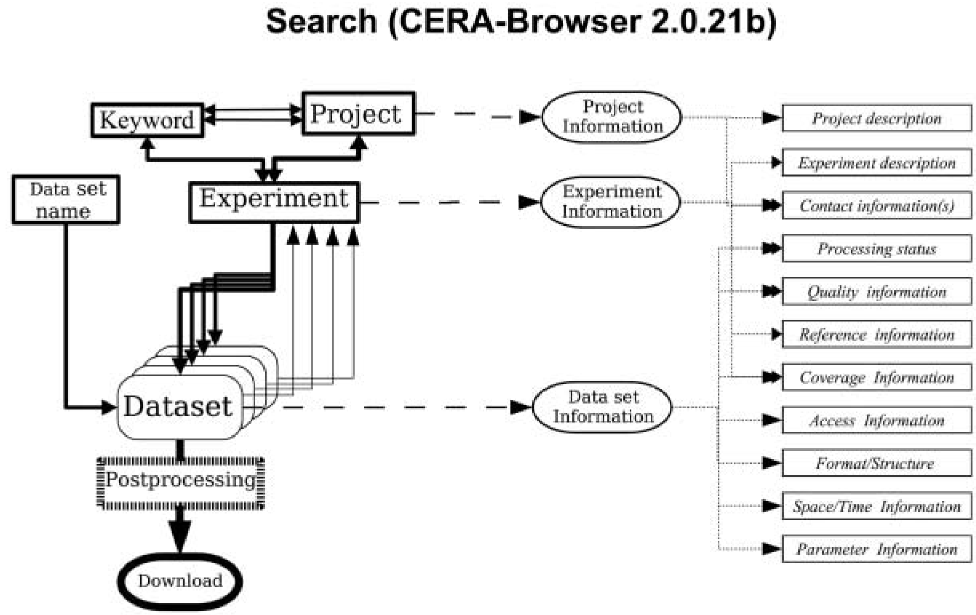

Fig. 5. Search path in the CERA-2 web interface. The search begins with the selection of a keyword or a project (e.g., CEOP). The related experiments contain the datasets that can be downloaded in full or in part.

5. Technical basis

5.1 Data warehouse

The vast majority of data at WDC-Climate are global model output of numerical climate experiments that were previously carried out using high-performance computers. These globally gridded values of physical parameters are stored as time series in WMO-GRIB format (see http://www.wmo.ch/web/www/WDM/Guides/Guide-binary-2.html) or in NetCDF. Furthermore, the WDC-Climate archives disseminate observational data sourced from various projects related to Earth System Research.

For a number of scientific projects, WDC-Climate is the central data-holding facility in the distribution of data between partners and global users during and after the project duration. Among such research projects are, for example, CARIBIC (Civil Aircraft for Regular Investigation of the Atmosphere, Based on an Instrumentation Container), COPS (Convective and Orographically-Induced Precipitation Study), and model data for scenarios provided by the Intergovernmental Panel on Climate Change (IPCC). For the latter, the WDC-Climate operates the Data Distribution Centre (DDC; http://ipcc.wdc-climate.de/). Furthermore, the results of reanalysis projects from the European Centre for Medium-Range Weather Forecasts (ECMWF) and NCAR are available for download, subject to the access conditions of the respective reanalysis centres.

The metadata of all database entries at WDC-Climate can be browsed anonymously using the web-based user interface. Direct access to the data-download area requires the user to log in to the system (see Section 4.2). Individual data-access rights are connected to each WDC-Climate user account.

5.2 Hardware at WDC-Climate

Four *Storage Tek STK* Silos form the DKRZ tape-based mass storage system that currently hosts more than 4 petabytes of data from different projects and climate research partners. The web-accessible component of the data at WDC-Climate is relatively small, making up approximately 250 terabytes of the data housed on the system.

The tape data are cached on 60 terabytes of disk; however, when some of the data of the ORACLE database management system have been migrated to tape, several minutes is required to mount the tape, copy the data to the disk cache, and make the data available online.

For this reason, CEOP MOLTS data tables, which are approximately 100 gigabytes, are kept online during the project to provide superior data-access performance for CEOP researchers.

The database itself runs as a distributed system on five NEC TX7 computers running Linux. Additional detailed information on the hardware used by WDC-Climate can be obtained from the DKRZ homepage (http://www.dkrz.de/).

5.3 Structure of the CERA-2 database

The core of WDC-Climate is the integrated data and metadata model CERA-2 (http://www.mad.zmaw.de/wdc-for-climate/cera-data-model/), which was first described by Lautenschlager et al. (1998). CERA-2 runs on an Oracle Relational Database Management System (RDBMS). Data housed in completely new and different structures can be integrated into its flexible schema and described in detail without changes in the data model, i.e., the table structure.

According to the CERA-2 data model, every entry for an experiment or dataset (see Section 4.2.) is described in a basic form via a group of tables. This description is connected to other groups of tables, so-called blocks (Fig. 6), each of which refers to a certain metadata topic such as a keyword list, contact information, or a description of spatial coverage. This structure can easily be adapted to new requirements. Additional table groups (CERA-2 Modules) describe data access, data structure, and geographical information. Thus, CERA-2 provides a comprehensive description of the stored information with respect to data description and usage.

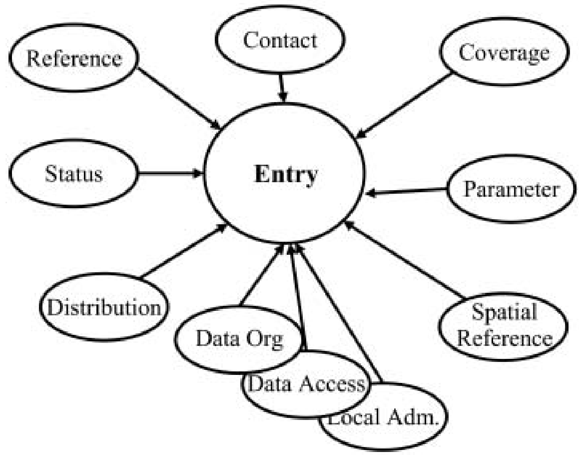

Fig. 6. In the CERA-2 data model, every entry (experiment or dataset) is connected to various specific metadata.

The metadata entries form a two-level hierarchy that consists of so-called experiments and the datasets related to the experiments. At the experiment level, the data are coarsely grouped according to the climate model and data type (e.g., grid vs. MOLTS). At the beginning of 2006, CEOP data comprised approximately 23 experiments in WDC-Climate that mainly reflect the original (raw) data structure provided by the NWP centres. The metadata for each experiment is split into various dataset descriptions that refer to the files submitted to the WDC-Climate. After restructuring the data, each dataset contains a time series of global longitude–latitude grids of one variable at one altitude level (two-dimensional), which is the most frequently used structure in climate applications.

5.4 Direct web access to the metadata catalogue

During the past decade, xml (extensible mark-up language) has become the most frequently used standard for the exchange of metadata between data storage systems. CERA-2 also provides xml metadata output for direct access to all information. This provides an alternative method of access to catalogue data. Triggered by an Internet request (a so-called http protocol), a servlet returns the metadata output of the required CERA-2 entry. This data can be mapped to html or any metadata format such as ISO, Dublin Core, or the metadata protocol of the Open Archive Initiative (OAI). Projects can use this service to integrate WDC-Climate catalogue metadata into local project websites or assimilate them into their own data catalogue.

5.5 Direct internet access to binary objects

An interactive means of establishing web access to WDC-Climate data is via the web browser (GUI, see Section 4.2) and downloading BLOB data to the local client disk; however, the Java programme jblob is a command-line interface that enables direct internet access to

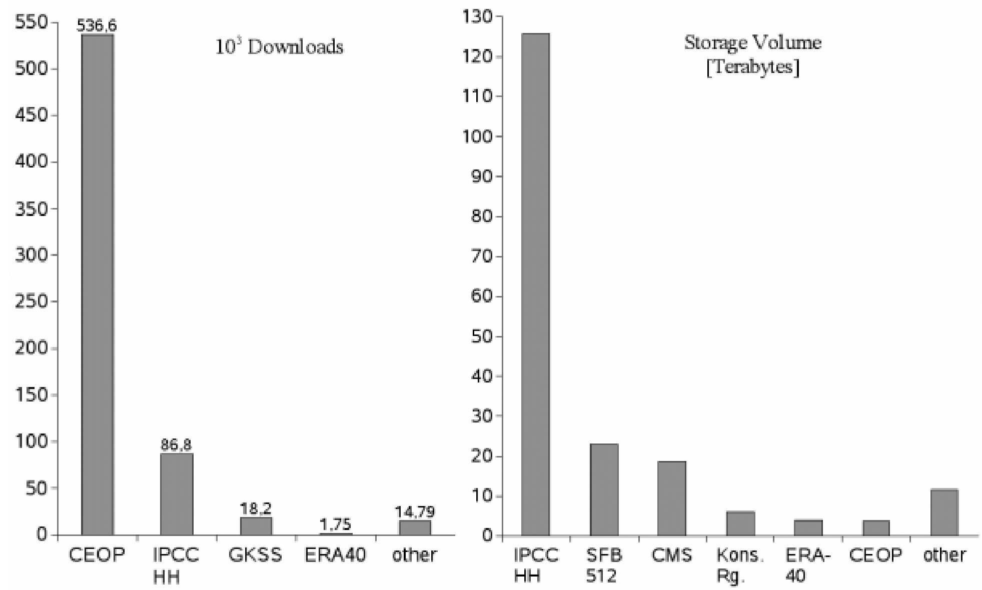

Fig. 7. In terms of data volume and the number of download requests, CEOP is one of the major projects at WDC-Climate. Left: Total number of WDC-Climate data downloads per research project in 2005 (10^3). Right: WDC-Climate data-storage volume per research project in December 2005 [terabytes].

the data, provided that the WDC-Climate entry identification can be specified. This is a useful tool for extended data downloads by various scripts.

5.6 Direct web access to MOLTS

A servlet has been implemented at WDC-Climate to provide direct access to the database tables that contain CEOP MOLTS as single real numbers. The start and end times and one or more variables must be specified in the URL (Uniform Resource Locator) request to obtain an ASCII time-series table for a certain reference site, based on the model calculation of one of the twelve CEOP model data providers.

This service is mainly used by WTF-CEOP, but it is now open to all members of the CEOP community. These data can be integrated into any local application, i.e., they can be accessed and downloaded by any client programme that has access to the World Wide Web.

6. CEOP—an integrated part of the WDC-Climate warehouse

CEOP model output is an integrated part of the WDC-Climate data warehouse that also provides IPCC climate simulations using the MPI-M climate model (ECHAM5) for Assessment Report 4, selected IPCC scenario data from other research units (IPCC-DDC), data from certain analysis projects, and other data that might also be of interest to the CEOP community.

In terms of the number of data downloads and storage fraction, the CEOP project is an important part of the WDC-Climate data warehouse (Fig. 7). Collaborations with various CEOP partners and feedback from users have led to an improved integration of all of the model output from the CEOP NWP centres. This provides a promising and stable basis for future CEOP activities.

The lessons learned from the CEOP Phase I data-management activities include the fact that the integration of model data into a central database archive requires greater coordination and clearer interface definitions than that achieved in the past. The huge amounts of data that are physically transferred around the globe require increasingly effective use of available but limited resources. The CEOP model output archive presently integrates more than 4 terabytes of data. This archive is dependent upon agreed interface definitions to import and homogenize model output, as well as a continuing focused discussion of the scientific applications that are likely to involve access and dissemination of the output. CEOP Phase II will need to place a greater emphasis on these latter points.

Acknowledgements

We wish to thank all of the CEOP NWP centres for their contributions to the CEOP model output archive. Special thanks are due to Lawrie Rikus (BMRC), Burkhardt Rockel, and Beate Geyer (GKSS-Forschungszentrum) for their support in converting the data structures and data formats, as well as the NWP centres for uploading their model output to WDC-Climate. Additional thanks are due to Michael G. Bosilovich for numerous fruitful discussions and suggestions. Finally, we especially acknowledge the comments and suggestions made by the reviewers and editors.

References

Lautenschlager, M., F. Toussaint, H. Thiemann, and M. Reinke, 1998: The CERA-2 Data Model, Deutsches Klimarechenzentrum, Technical Report No. 15, Hamburg, July, www.dkrz.de.

All URLs in the previous text were active as of January 2006.