Data Analysis System Attached to the CEOP Centralized Data Archive System

Toshihiro NEMOTO

Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

Toshio KOIKE

Department of Civil Engineering, The University of Tokyo, Tokyo, Japan

and

Masaru KITSUREGAWA

Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

(Manuscript received 8 March 2006, in final form 19 December 2006)

Abstract

A large amount of data is being collected and archived in the Coordinated Enhanced Observing Period (CEOP) project to help to increase researchers' understanding and knowledge of the global water cycle system. We describe the data archive system that integrates data (in-situ, satellite, and model output) produced by the CEOP project. We also detail the user interface for analyzing the data. First, we explain the characteristics of the global water cycle data to be archived and the functions requested by users to analyze the data. We then explain the architecture of the archiving system and its graphical user interface (GUI) that we are currently constructing. The interface will integrate data of various dimensions, temporal and spatial resolutions, coordinates, precision, and format. This interface will provide users with the environment to handle data irrespective of type.

1. Introduction

Worldwide, there are many issues associated with water including problems such as water shortages, heavy rains, pollution, and the destruction of ecosystems. Food shortages and infectious diseases caused by these water associated problems are occurring more frequently, especially in developing countries. Such problems are caused by fluctuations in the water cycle and by social factors such as the rapidly increasing demand for water because of overpopulation, urban development, and industrial growth. Understanding these water cycle fluctuations and improving the precision of forecasting are crucial to resolving the water crisis. Under such conditions, GEWEX (Global Energy and Water Cycle Experiment) has initiated the CEOP (Coordinated Enhanced Observing Period) project, which has since October 2002 collected and archived a huge amount of data. There are three kinds of CEOP data: in-situ, satellite, and model output, all with different dimensions, spatial and temporal resolutions, precision, formats, and coordinate systems. A system to integrate these data and to improve understanding and prediction is needed.

Corresponding author: Toshihiro Nemoto, Institute of Industrial Science, University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo, 153-8505, Japan.
E-mail: nemoto@tkl.iis.u-tokyo.ac.jp

In this paper, we explain the centralized CEOP data archive system that we are currently constructing in order to make it more available and easier to use. The system's data server manages all of the data, including metadata. Though the CEOP project has three data archive centers for three different kinds of data, respectively, the centralized data archive system replicates all of the CEOP data in the other data archive centers and stores them. Though the server uses a tape library system and disk arrays for storage, the location of the data is hidden from users, and so users can retrieve data without considering its location. The server provides users with a menu-based, integrated graphical user interface for data retrieval and analysis. Users can access all kinds of data through the same interface without taking account of data type. Depending on its dimensions, users can view the retrieved data as graphic charts or bitmap images. Some analysis operations such as average, difference, correlation, and so on can be applied to one or more retrieved data items on the server through the graphical user interface (GUI). The preliminary data archive system has already been implemented and is now being used experimentally.

First, we summarize the requirements for archiving and analyzing CEOP data. Then, we explain the data to be archived and the analysis functions. Finally, we describe the design and architecture of the data system in detail and introduce the system's user interface.

2. Requirements for data archive system

2.1 Archived data

There are three kinds of data archived by the CEOP project. They are in-situ, model output, and satellite data. In addition to these data, their metadata also must be archived.

a In-situ data

In-situ data are temporal series of the values observed at 35 reference sites around the world. Each reference site has one or more stations. The in-situ data are divided into three categories, namely surface, subsurface, and upper air observation. Surface observation includes air temperature, pressure, humidity, precipitation, heat flux, and radiation at the ground level. Subsurface observation includes soil temperature, soil water content, and soil heat flux below the earth's surface. Upper air observation includes air temperature, humidity, and pressure as measured by radiosonde. Though the sorts of values to be observed and their sensor heights are recommended by the CEOP project, not all values are always observed at each reference site. The observed values, therefore, depend on the reference site: some reference sites have more than two stations to observe one value, and some observe one value at multiple heights. Frequency of observation also varies between reference sites.

Each reference site checks the quality of all observed values, assigns them quality flags, and converts them into CEOP standard format files. The National Center for Atmospheric Research (NCAR) at Boulder, USA collects all values from the reference sites, performs set of quality assurance procedures, and archives and distributes them. The total amount of in-situ data for two years and three months is almost 600 MB.

b Model output data

Model output data are gridded values generated from global forecast models or assimilation systems at 11 weather forecast centers. Two types of model output, namely gridded data and a site-specific time series from each of the reference sites, are archived.

The gridded data are two- or three-dimensional data. Each cell has several prognostic variables, such as air temperature, humidity, and pressure. Though the sorts of values to be output are defined by the CEOP project, some forecast centers do not output all values because of their weather analysis models. The coordinate system and the resolutions in the time and space axes are different among the forecast centers. For example, some forecast centers use a latitude-longitude grid system and other forecast centers use a Gaussian grid system in space axes. Pressure level, model layer, and so on are represented in the vertical axis. The several weather analysis processes for forecasting are executed in parallel at intervals of several hours. Accordingly, there are more than two values corresponding to each single physical variable and each single temporal point. Each has its own unique reference time, which is the moment the weather

analysis started. Each forecast center converts its model output data to GRIB format and sends them to the Max-Planck Institute at Hamburg, Germany, which then makes them publicly available. The total amount of model output data is almost 20 TB per year.

The site-specific time series values are designated as Model Output Location Time Series (MOLTS). They are one-dimensional time series of variables at 41 points including the 35 reference sites extracted from gridded data. Because the MOLTS data are generated in order to compare the corresponding in-situ data in detail, they have higher temporal and vertical resolution in some weather forecast centers than the gridded data does. The sorts and number of the variables and the horizontal coordinate system and resolution differ among the forecast centers. The native file formats are also different among the forecast centers. Recently, the CEOP Data Management group and model data providers have agreed to adopt the Climate and Forecast (CF) compliant NetCDF format for the CEOP model output data. The work on data conversion is currently on-going.

c Satellite data

The satellite data are remotely sensed data from satellites sensors, such as DMSP SSM/I, TRMM TMI, TRMM PR, GMS S-VISSR, NOAA AVHRR, TERRA/AQUA MODIS, and AQUA AMSR-E, which have been operating during the CEOP period. Satellite data suppliers such as NASA, ESA, and JAXA process the satellite data. They collect data radiometrically and geometrically and resample it into maps of three sizes: maps of the regions around the 35 reference sites, maps of five monsoon regions, and a global map. The satellite data have a latitude-longitude grid and their resolution is almost the same as the original resolution of the sensors. Resolution therefore, depends on the satellites and the sensors. Though the variables of the satellite data in the CEOP project depend on the sensors, the satellite data include both lower level products such as brightness temperature and albedo and high level products such as sea surface temperature, soil moisture, and so on. Most of the satellite data are two-dimensional data from the earth's surface. Some sensors can observe the vertical distribution of physical values and, in this case, the data are three-dimensional. The data formats are different among the data suppliers. The total amount of satellite data amounts to more than 100 TB per year.

d Metadata

The CEOP committee has designed a standard format for satellite data that is based on ISO 19115. Some items were added to it to fill the gap between ISO 19115 metadata items and user requests. The metadata files for satellite imagery data are written in XML and provided together with the satellite imagery data files. The CEOP committee is currently discussing standard formats for the metadata of model output and in-situ data, but they have not yet been finalized. However, the information necessary to use model output and in-situ data is written in various non-unified formats and can be obtained from the data archive centers.

2.2 Functions to be implemented

Archived data are used for investigation of meteorological phenomena, improvement of numerical meteorological forecast simulation precision, refinement of the processing methods for satellite data, and so on. The functions required for these purposes vary with the analysis needs of the users. We discussed with some scientists the functions to be implemented. The functions, most of which have already been implemented to the CEOP centralized data archive system, are summarized below.

a Data retrieval from the archive

CEOP data are identified by at least three kinds of information, namely type of data, date and time, and area. The type of data is specified by precise variable name such as air temperature, pressure, and soil moisture in addition to category name such as in-situ, satellite, and model output data. The area is specified by the name of the point or the region such as the particular reference site, the monsoon region, or global data. Therefore, data can be retrieved by condition about the type of data, date and time, and the area. Furthermore, in order to specify the requested data more precisely, it is desirable for users to describe the spatial and temporal coordinate system, the units and so on. In addition, as the sorts of observed param-

eters in in-situ data depend on the reference sites and the sorts of provided variables in model output data differ between forecast centers, users can send search queries such as 'what kind of air temperature data is available at this particular point'.

In the centralized data archive system, data are retrieved from the archive by specifying three items; data type, geographical location and time period. Data specification is based on menus and all available items are shown to the users.

b Format conversion

CEOP data have various file formats. Though the in-situ data have a unified format, it is only standard within the CEOP project. Currently, the MOLTS data have different file formats among the forecast centers. Therefore, when the users process the CEOP data with their application programs, it is necessary to convert the data into a file format the application programs can access. Even when users process them using a user–created program, it is useful to convert the file format. When the file format is unified, only the converted format must be supported by the program. Accordingly, it is necessary to convert the format to one popular in the meteorology and oceanography field, such as NetCDF or GRIB.

Through the client program of the centralized data archive system, users can save the retrieved CEOP data in ASCII text or NetCDF formats.

c Alignment and adjustment of temporal and spatial axes

Since the resolution and the coordinate systems differ by data in the CEOP project, it is not possible to compare them directly with each other. To compare one data set with another, the resolution and coordinates of all axes in the data must be standardized.

The common methods of arranging the axes are resampling with the nearest neighbor method, linear interpolation, and spline interpolation. These methods can be applied to the global water cycle data. In addition to these methods, methods peculiar to the characteristics of the physical values are also necessary. For example, to convert hourly accumulative values to daily ones, 24 hourly values should be added. To convert the highest temperature and the lowest temperature, maximum and minimum values from the values in the corresponding region should be selected. For average values, it may be necessary to recalculate the values with adequate weights. Because a suitable alignment and adjustment method depends on both the data characteristics and the user's analysis needs, the user must be able to select the method to be applied.

The alignment and adjustment methods supported in the centralized data archive system are the nearest neighbor method, linear interpolation, summation, and maximum, minimum, and average values.

d Aggregation

Aggregation operations such as averaging, maximizing, minimizing, variance calculation, summing, and so on are also required. In addition to aggregation values in a single temporal and spatial region, for example the calculation of average values in a 250 km $\times$ 250 km area around the reference site, interval aggregation values in every regional unit, such as the maximum values for every day, are important. Moreover, diurnal, seasonal, latitudinal, and longitudinal aggregation, as well as aggregation of multiple arbitrary points or areas, aggregation of only land or sea area, and so on are also necessary.

The centralized archive system supports averaging, maximizing, minimizing, variance calculations, summing in single and multiple axes, and interval, diurnal, latitudinal and longitudinal aggregation. As well, masking operations, which makes it possible to aggregate in more complicated regions, are supported in the centralized archive system.

e Arithmetical calculation and analysis operation

To process data analysis, various arithmetical calculation operations are necessary. For example, deriving new variables from other variables, comparison of differences between two variables, correlation coefficients, regression coefficients, temporal differential calculus of sequential values, rotation and divergence of more than two dimensional data, and so on are required. It is necessary to apply these operations recursively.

The frequently applied operations must also be easily invoked. Users are demanding the

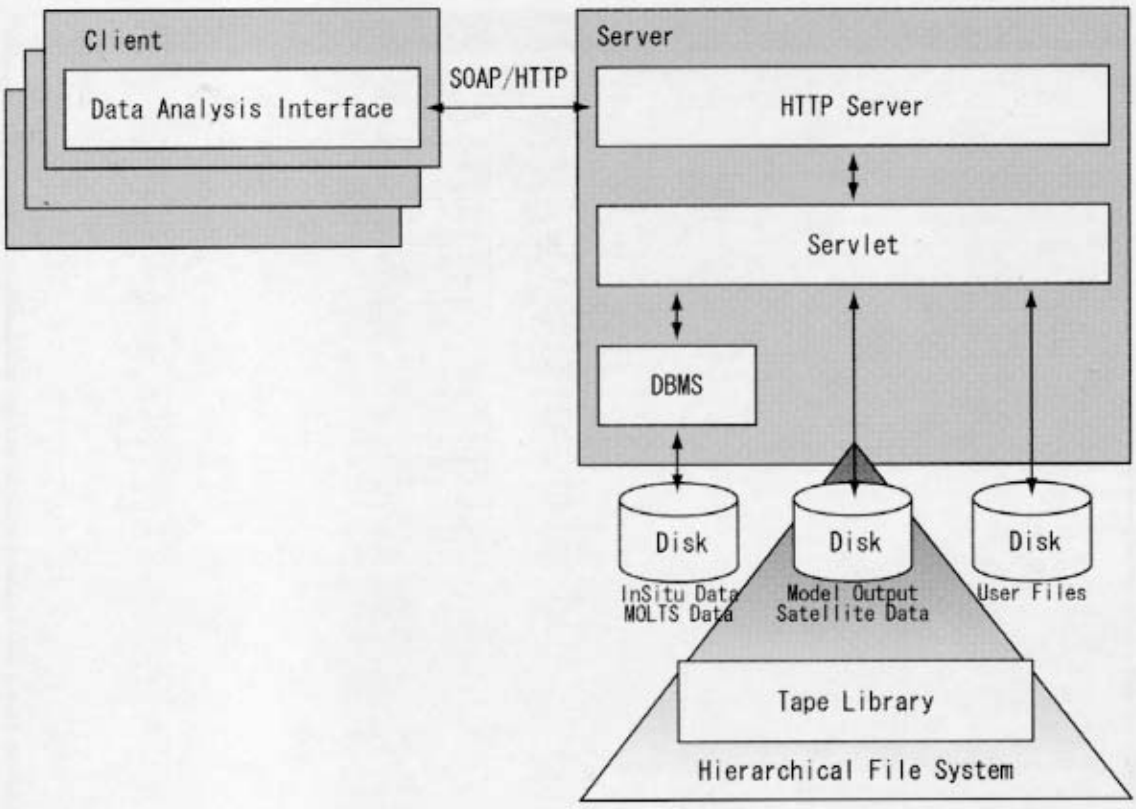


Fig. 1. System architecture of the data archive system.

ability to define and register frequently applied operations, as these depend on the users and the analysis methods.

The centralized archive system can easily derive correlation and regression coefficients. Moreover, the system can calculate arithmetic equations including some fundamental user specified ones.

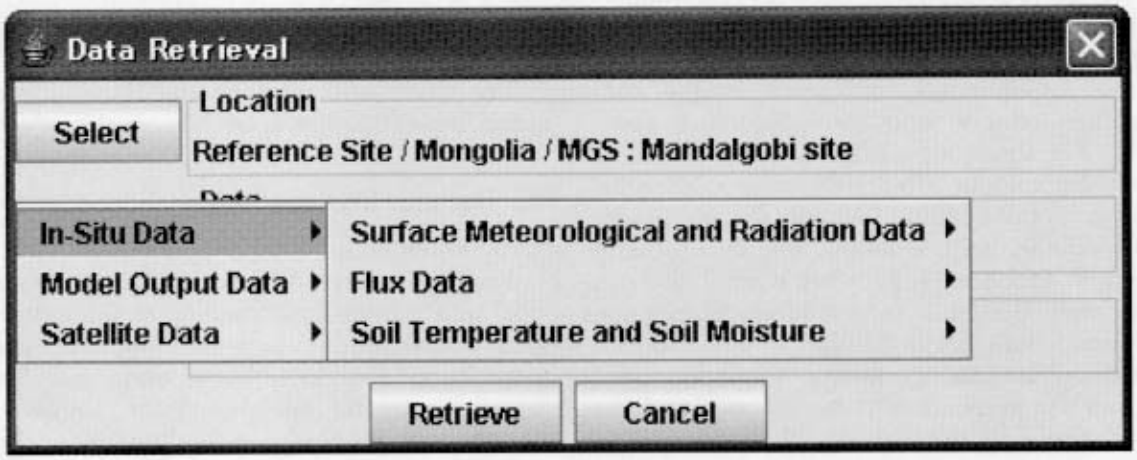
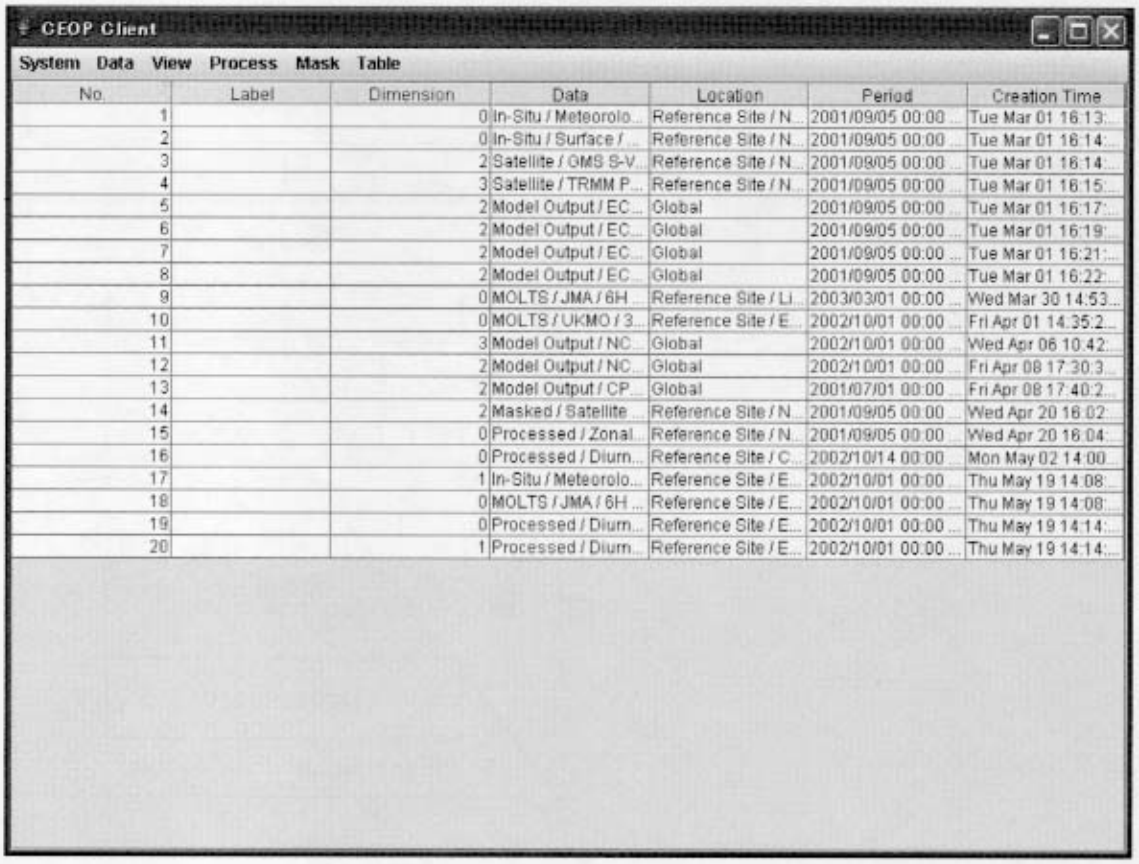


Fig. 2. Parameter input window for data retrieval (Second layer in the menu for data type selection.)



No.	Label	Dimension	Data	Location	Period	Creation Time
1		0	In-Situ / Meteorolo...	Reference Site / N...	2001/09/05 00:00 ...	Tue Mar 01 16:13:...
2		0	In-Situ / Surface /...	Reference Site / N...	2001/09/05 00:00 ...	Tue Mar 01 16:14:...
3		2	Satellite / GMS S-V...	Reference Site / N...	2001/09/05 00:00 ...	Tue Mar 01 16:14:...
4		3	Satellite / TRMM P...	Reference Site / N...	2001/09/05 00:00 ...	Tue Mar 01 16:15:...
5		2	Model Output / EC...	Global	2001/09/05 00:00 ...	Tue Mar 01 16:17:...
6		2	Model Output / EC...	Global	2001/09/05 00:00 ...	Tue Mar 01 16:19:...
7		2	Model Output / EC...	Global	2001/09/05 00:00 ...	Tue Mar 01 16:21:...
8		2	Model Output / EC...	Global	2001/09/05 00:00 ...	Tue Mar 01 16:22:...
9		0	MOLTS / JMA / 6H ...	Reference Site / Li...	2003/03/01 00:00 ...	Wed Mar 30 14:53...
10		0	MOLTS / UKMO / 3...	Reference Site / E...	2002/10/01 00:00 ...	Fri Apr 01 14:35:2...
11		3	Model Output / NC...	Global	2002/10/01 00:00 ...	Wed Apr 06 10:42:...
12		2	Model Output / NC...	Global	2002/10/01 00:00 ...	Fri Apr 08 17:30:3...
13		2	Model Output / CP...	Global	2001/07/01 00:00 ...	Fri Apr 08 17:40:2...
14		2	Masked / Satellite ...	Reference Site / N...	2001/09/05 00:00 ...	Wed Apr 20 16:02:...
15		0	Processed / Zonal...	Reference Site / N...	2001/09/05 00:00 ...	Wed Apr 20 16:04:...
16		0	Processed / Diurn...	Reference Site / C...	2002/10/14 00:00 ...	Mon May 02 14:00:...
17		1	In-Situ / Meteorolo...	Reference Site / E...	2002/10/01 00:00 ...	Thu May 19 14:08:...
18		0	MOLTS / JMA / 6H ...	Reference Site / E...	2002/10/01 00:00 ...	Thu May 19 14:08:...
19		0	Processed / Diurn...	Reference Site / E...	2002/10/01 00:00 ...	Thu May 19 14:14:...
20		1	Processed / Diurn...	Reference Site / E...	2002/10/01 00:00 ...	Thu May 19 14:14:...

Fig. 3. Retrieved data management window (20 retrieved data items are listed with information about data type, location, and time.)

f Visualization

Data visualization is indispensable for data analysis. In the CEOP project, since many variables are provided, it is necessary to support many visualization methods suitable for each variable, for example, two-dimensional line charts, two-dimensional bitmap images, contour line images, and three-dimensional images. For these visualization methods, users must be able to arbitrarily assign temporal axis, horizontal spatial (latitude and longitude) and vertical axes (elevation), and data value to the x, y, and z axes. In other words, methods that show the cross section of four-dimension values of data perpendicular to any axis are necessary. It is also necessary to show the cross section not perpendicular to any axes and the cross section by the curved surface specified by the users. There is also demand for images that can be animated by moving the cross sections and, to compare more than two variables, for overlaying multiple lines or images. Indicating the corresponding points in multiple charts and storing the visualized images in the format and quality specified by the user are also important.

The client can visualize the retrieved data using several methods. Details are described in Section 3.5.

3. Data archive system

3.1 Design policy

The key concept of the data system is easy use. In a conventional system, to integrate in-situ, model output, and satellite data, users must carry out many bothersome operations before beginning analysis. First, they must search and retrieve appropriate data. Then they must convert them to a file format appropriate to the tool they use. Then, they may

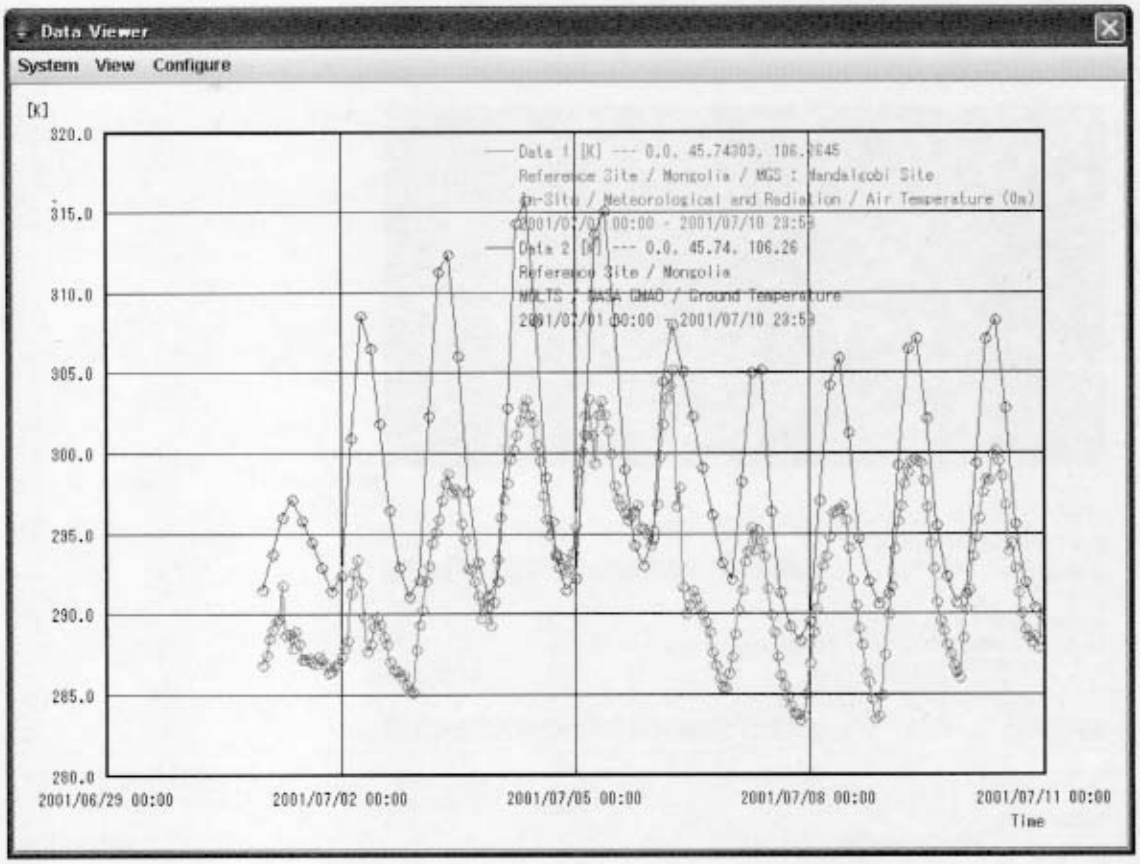


Fig. 4. Example of a line graph (Observed and actual air temperature in MOLTS by NASA GMAO (Global Modeling and Assimilation Office) at the Mongolia site between July 1st and July 10th in 2001.)

need to reproject, regrid, and subsample the data to reconcile their projection, coordinates, and area of interest. They may also need to interpolate the data to standardize their resolution. In addition to these spatial arrangements, users may need to rectify the data to synchronize the temporal resolution. After these operations, users can compare the different types of data. This preprocessing is unavoidable and laborious. The scientists working on global environmental data in particular are not always skillful at handling a large amount of data. Therefore, we designed the system to ensure that scientists could easily access and analyze data without special knowledge about the database or computer systems. In addition, CEOP data users are scattered all over the world, and their computer systems vary, so avoiding the need for special software and hardware is also important. It is necessary to support as many users as possible to improve the usefulness of the data.

3.2 *Internal data representation*

Though each data set in the CEOP project has its own dimension, all the data retrieved by the users are handled as unified internal data representation in the data system. All the data retrieved by the users are represented as a four dimensional array. The four axes are temporal, vertical, latitudinal and longitudinal axes. For example, the sequential values of air temperature at a reference site are represented as $n \times 1 \times 1 \times 1$ array values; they have n size on the time axis and 1 size in the vertical, latitudinal, and longitudinal axes. The retrieved data in the internal representation also have the information about themselves and four

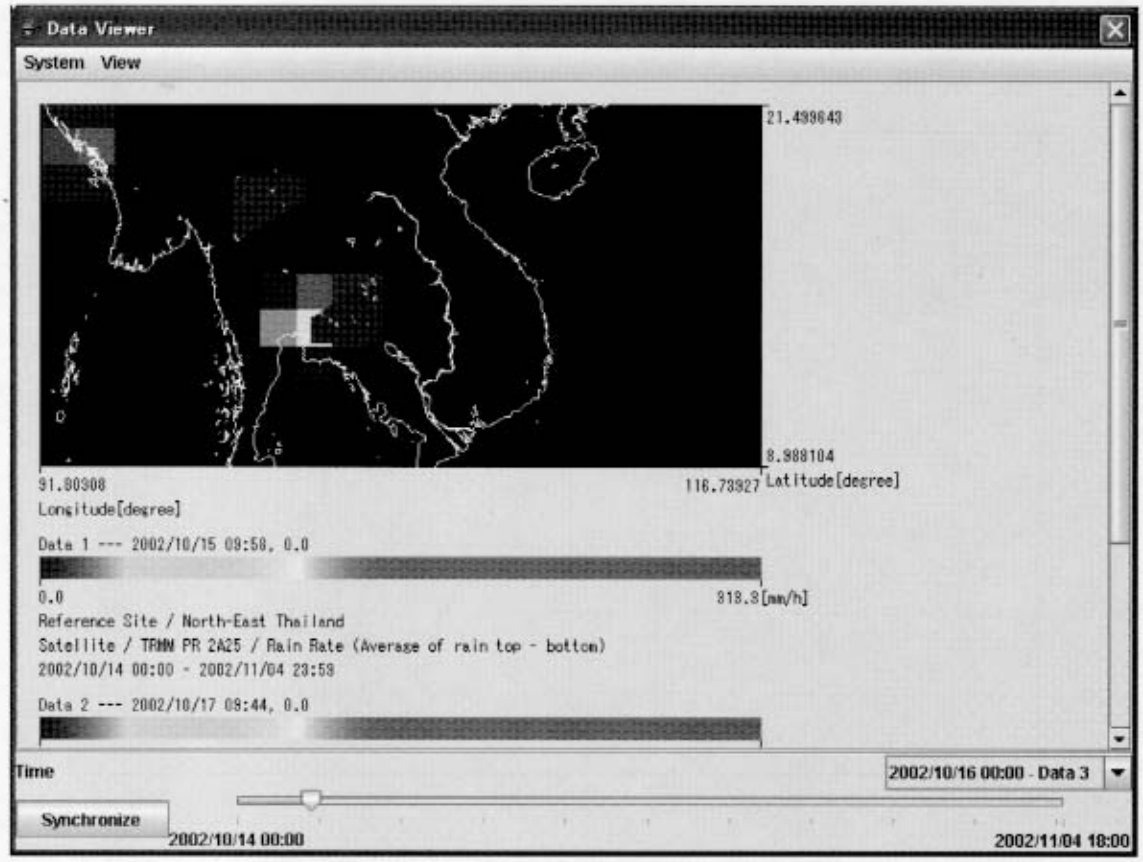


Fig. 5. Example of a bitmap image (Precipitation rate forecasted by the Japan Meteorological Agency and rain rate images around the North-East Thailand site and the Chao-Phraya River site by TRMM are overlaid.)

arrays of graduation values of all axes with description. In addition to its value, each cell in the retrieved four-dimensional array has a data quality flag. The information about retrieved data includes type, unit, conditions at retrieval, time of retrieval, and so on. The information about the axes includes the type of axis, unit of graduation values, and so on. The graduation values on each axis can represent both points and regions because some physical values such as sums or averages correspond not to points but to regions.

Using the unified internal data representation enables the archive system to uniformly manage three kinds of data and to offer integration functions. All retrieved data are represented with the same structure in the archive system, even if their original data types, dimensions and resolutions differ. This makes it possible to adjust and easily compare different kinds of data. The internal format is hidden from the users so that they do not need to consider the internal representation of data during their analyses.

3.3 Architecture of data archive system

The data archive system is based on a client-server model. The architecture of the system is shown in Fig. 1. The communication protocol between the server and client is a simple object access protocol (SOAP) over HTTP. HTTP is not always suitable for data transfer, but is widely used and, therefore, may cause few troubles, especially concerning the firewall. As SOAP is an XML-based standard protocol for exchanging objects, users' programs can directly access the data server. Therefore, we adopted SOAP over HTTP as the communica-

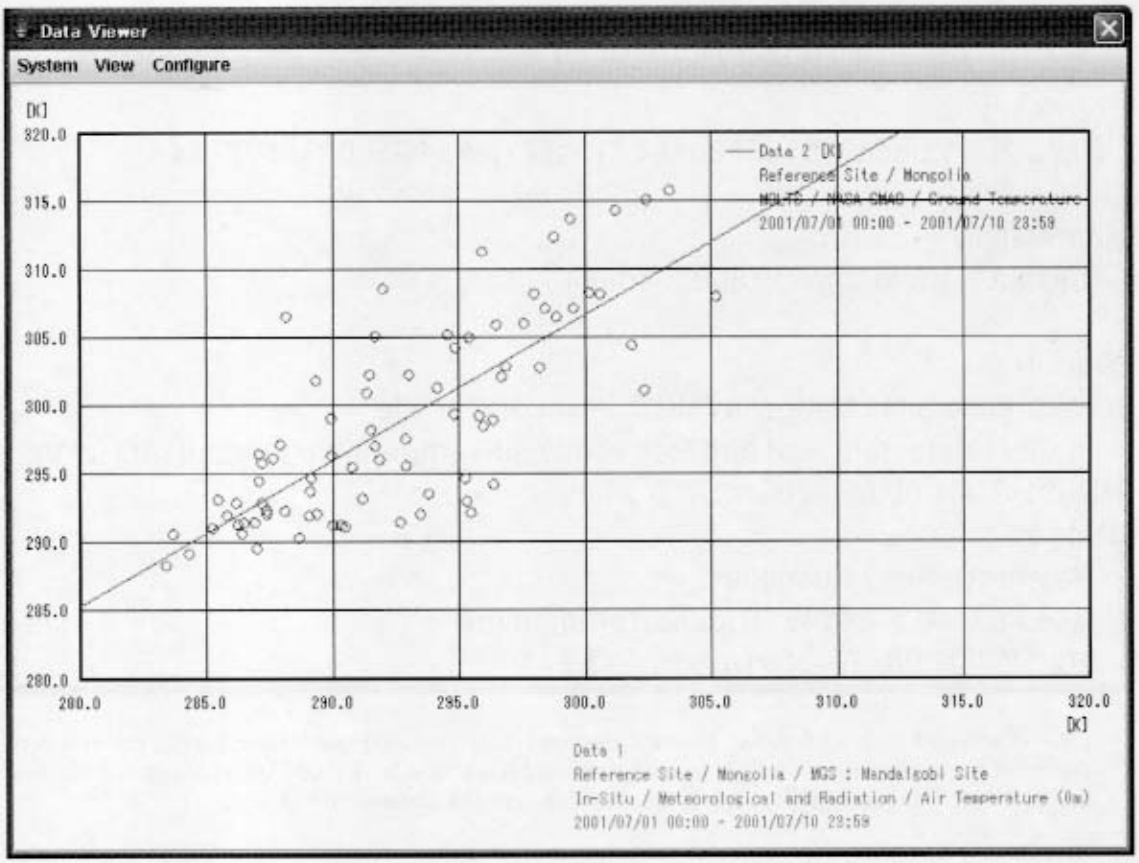


Fig. 6. Example of scattering graph (x axis represents observed air temperature at the Mongolia site between July 1st and July 10th in 2001 and y axis represents air temperature in MOLTS by NASA GMAO at same site and for the same period; target values are the same as in Fig. 4.)

tion protocol. However, as the amount of water cycle data is huge, representing all values in XML is not efficient. Therefore, to reduce the time required to transfer data between the server and the clients, all values to be transferred are represented as IEEE754 binaries and compressed. The compression method is lossless and compressed data are automatically extended when the server and clients receive them; therefore, users do not have to worry about the compression system.

3.4 *Data server*

The data server manages user accounts, retrieves values from storage, and provides clients with conversion of temporal and spatial axes, calculation and aggregation operations, and so on.

One-dimensional data such as in-situ data and MOLTS data are stored in a database management system (DBMS), but two- and three-dimensional data such as gridded model output and satellite data are stored as files on the hierarchical file system and only their metadata are stored in DBMS. There are several reasons why we do not manage two- or three-dimensional data as Large Objects (LOB) in DBMS. First, accessing LOB in DBMS is slower than accessing a file (Stolte et al. 2003). Second, existing implementations of LOBs tend to lack support for the hierarchical storage management system. Although the gridded model output data and the satellite data are stored on the hierarchical file system, small images around the reference sites clipped from the global data are stored on disks. The gridded model output data and the satellite data are too large to be stored on disks, and most users

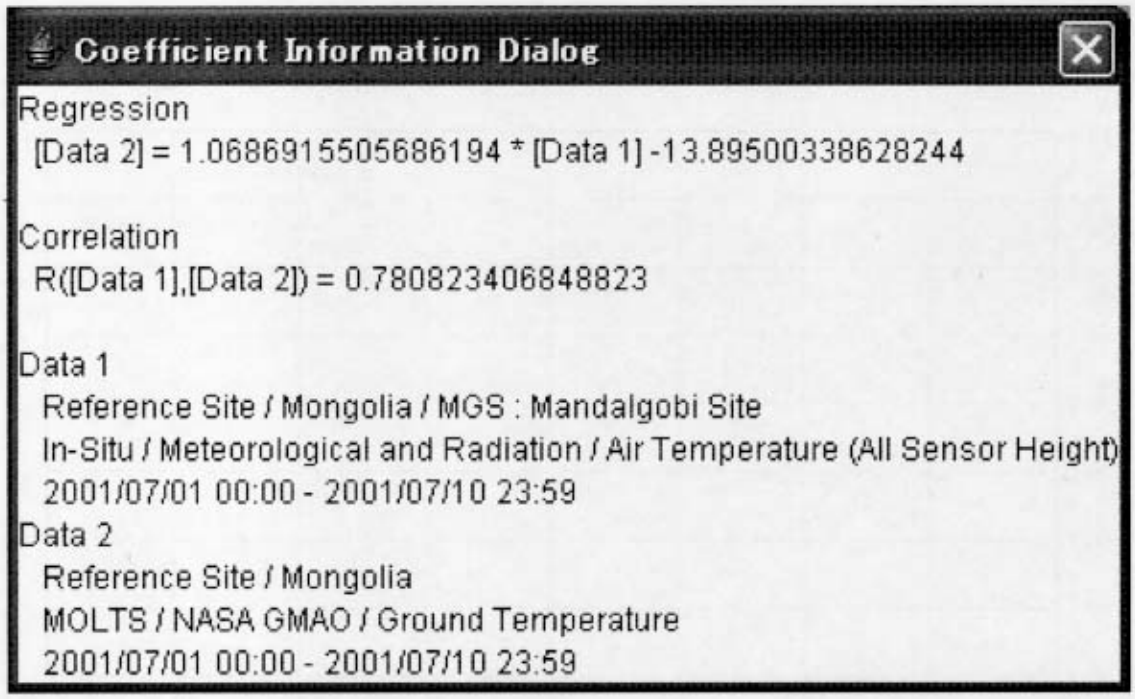


Fig. 7. Coefficient information window (Regression coefficients and correlation coefficient between observed air temperature and air temperature in MOLTS by NASA GMAO at Mongolia site between July 1st and July 10th in 2001; target values are the same as in Fig. 4.)

do not need global data. Generally, scientists need only the values around the reference sites to compare the values from ground observations and those of model output or of remotely sensed data. Therefore, storing these small portions on disks instead of in a hierarchical file system reduces response time. The location of the data is hidden from users, who do not need to consider where the data are stored. The data server automatically migrates and retrieves the appropriate data from DBMS, disks, or the hierarchical file system as the user requests and sends it to the clients. The DBMS manages the one-dimensional data and the metadata for all data sets. We use a commercial DBMS and JAVA database connectivity (JDBC) for the connection between DBMS and the data manager.

The data server is a servlet program. It receives requests from the clients through the HTTP server and then generates structured query language (SQL) commands for data search or executes analysis operations according to user requests. When a user sends data retrieval requests to the server through the user interface, the data server extracts a portion from the DBMS or the files, converts it to internal representation, and stores it in the user area in the server. The user can apply operation methods to the extracted data in the user area. The results of the operation method are also stored on the user area as a new file in the internal representation. Therefore, the user can apply the operation repeatedly without considering the difference between the retrieved data and the results of operations.

3.5 User interface for retrieval and analysis

The graphical user interface is the client program running in the user's computer. As it is written in JAVA, the client does not need any special hardware or software. Only a JAVA runtime environment is required. Since many kinds of current computers and operating systems support a JAVA runtime environment, the GUI for the data server works on many kinds of computers.

When the user starts the GUI program, an authentication window is shown. After passing

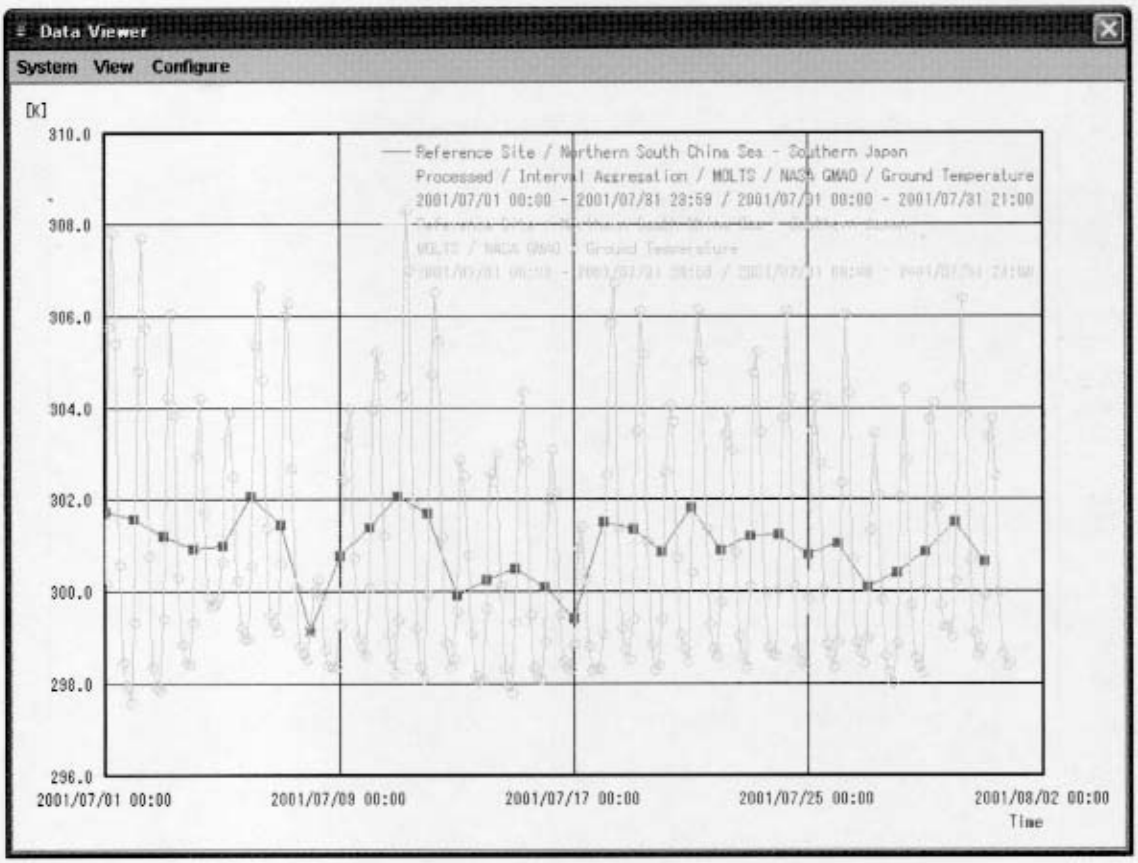


Fig. 8. Original values and results of temporal averaging (Ground temperature every three hours by NASA GMAO at the Northern South China Sea—Southern Japan site in July 2001 and derived daily average of ground temperature.)

the password check, the user can access the archived data. The requested data is specified by three items: data type, geographical location, and time period. The available data type, geographical location, and time period are listed in the menus and the user selects one of them (Fig. 2). The temporal period can also be specified by start date and time. After the request is transferred to the server, the server parses the requests, generates the SQL command, sends it to the DBMS, and stores the result in the user's area. The results are listed in the data manager window (Fig. 3) and are regarded as targets of analysis operations by the data analysis interface. A small amount of information about the retrieved data is displayed in the data manager window. Through the window the user sends a command to display more detailed information, to delete the data, to execute a calculation or aggregation operation, or to visualize the data. The user first selects one or more items in the list and then selects the command on the menu in the manager window. The user interface sends the appropriate request to the data server, and the data server executes the operation for the selected data in the user's area.

Figure 4 is an example of a line chart in which two data are drawn. The user can specify the temporal, vertical, latitudinal, longitudinal axis or data value as x axis and as y axis when a line chart is drawn. In Fig. 4, the time axis is the x axis and the data value is the y axis because the data selected for display are in-situ and MOLTS, which are time sequence values from a single point.

Figure 5 is an example of a two-dimensional bitmap image. In this image, model output

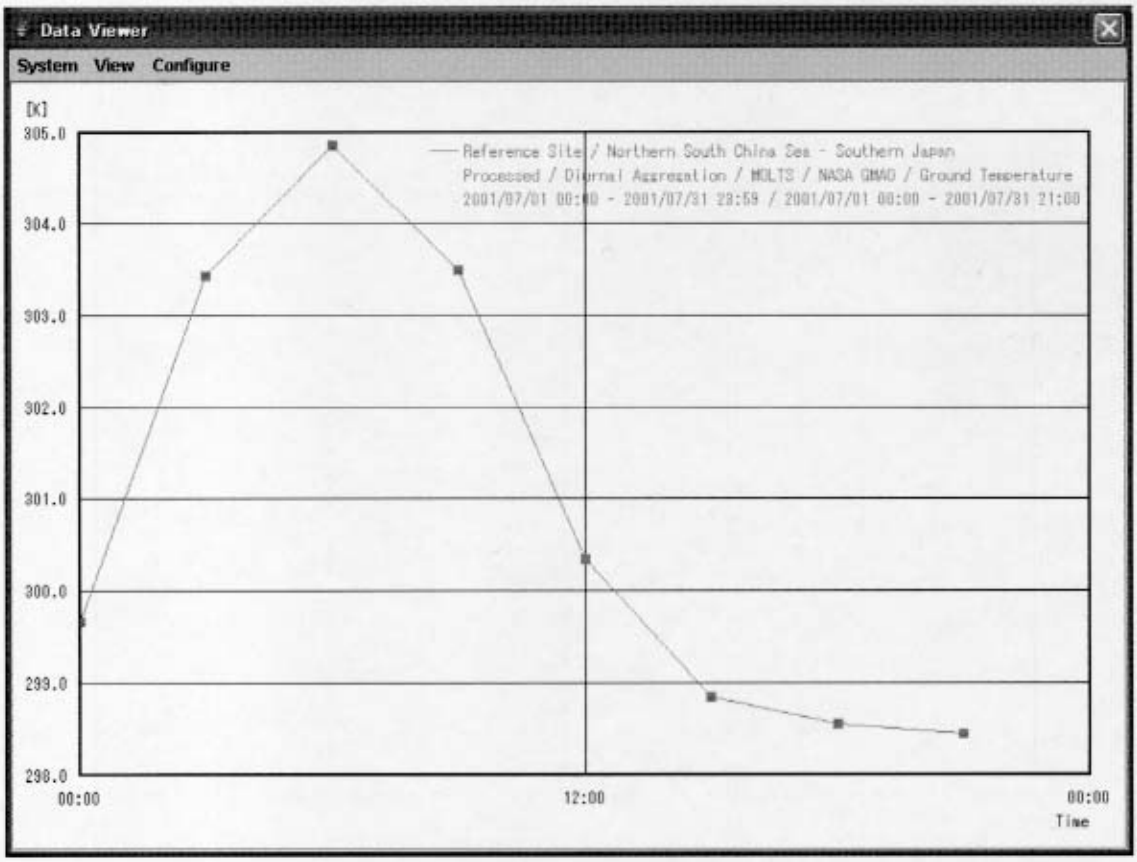


Fig. 9. Results for diurnal averaging (Diurnal average of ground temperature in MOLTS by NASA GMAO at the Northern South China Sea—Southern Japan site in July 2001; target values are the same as in Fig. 8.)

and satellite data are overlaid. To generate a bitmap image, the user can specify coloring method (color gradation or gray scale) and select whether contour lines, coastline, and point names are drawn. In addition to the line chart, the user can specify the temporal, vertical, latitudinal or longitudinal axis as the x axis and as the y axis when a line chart is drawn. In Fig. 5, the longitudinal axis is the x axis and the latitudinal axis is the y axis. The data selected to be displayed as a bitmap image are a time series of values, that is, the length of time axis is more than two. In this kind of case, where the length of the axis in the data that is not specified as an axis in the chart and the bitmap image is more than two, the slide bar is displayed below the chart or the image. The user can animate it by moving the slider. In addition, the user can synchronize multiple charts and images. To display charts and images, the user sends the requests to the server and the data values in the internal representation are transferred to the client. Then the client generates charts and images. If the server generates the charts and the images and sends them to the client, the charts and the images must be transferred whenever the user changes parameters for generating the charts and images. In addition, to display the numerical values of the variables on the charts and the images, the client needs the numerical values themselves in addition to the charts and images. This is the reason we enable the client to generate charts and images.

Figure 6 is an example of a scattering chart of two data. In Fig. 7 the correlation and the regression coefficients for the data are displayed. Scattering charts such as Fig. 6 are displayed

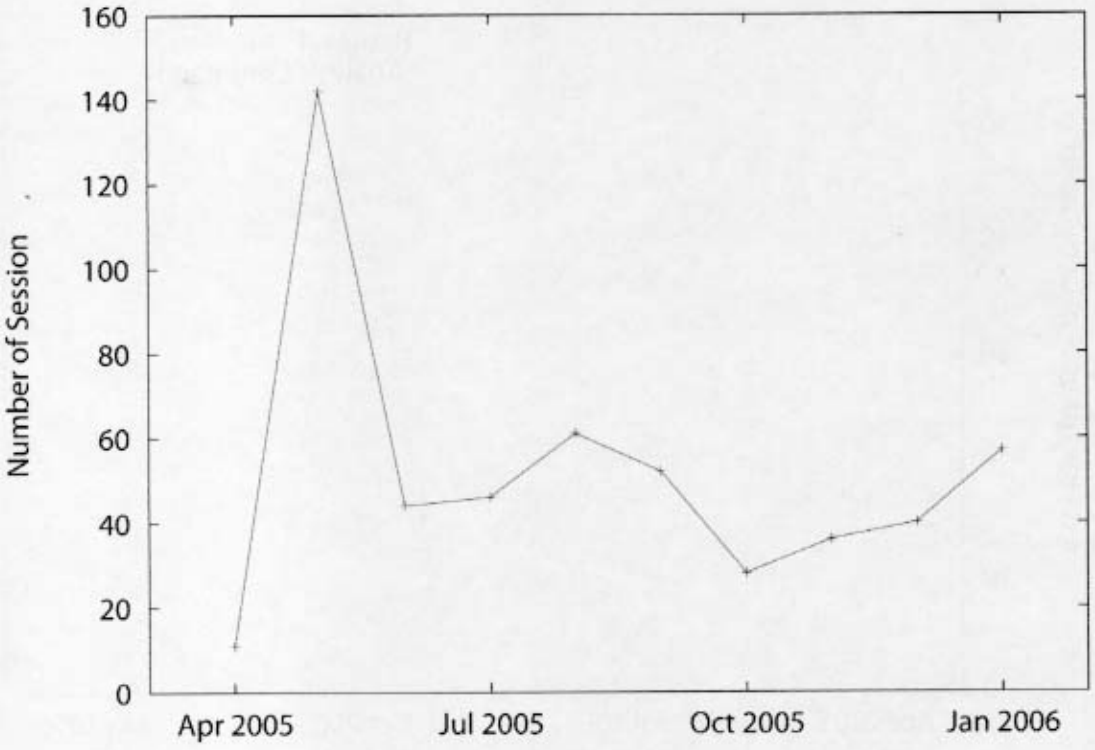


Fig. 10. Number of sessions to data archiving and analysis system from users.

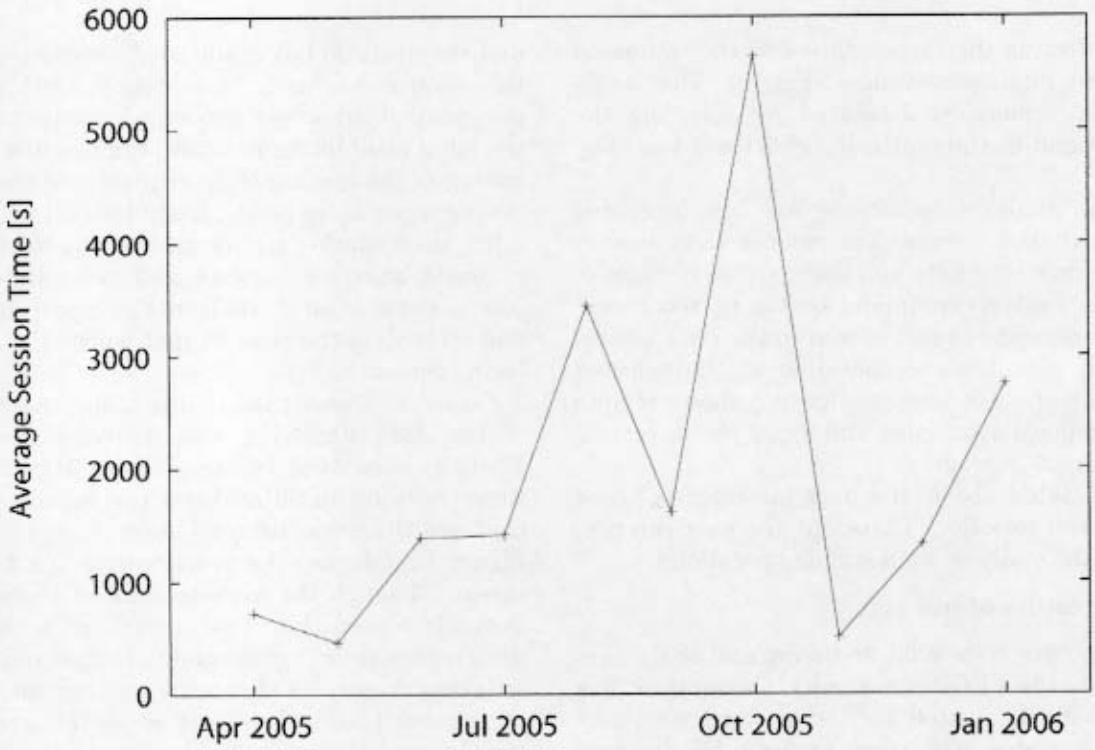


Fig. 11. Transition of average time of a single session.

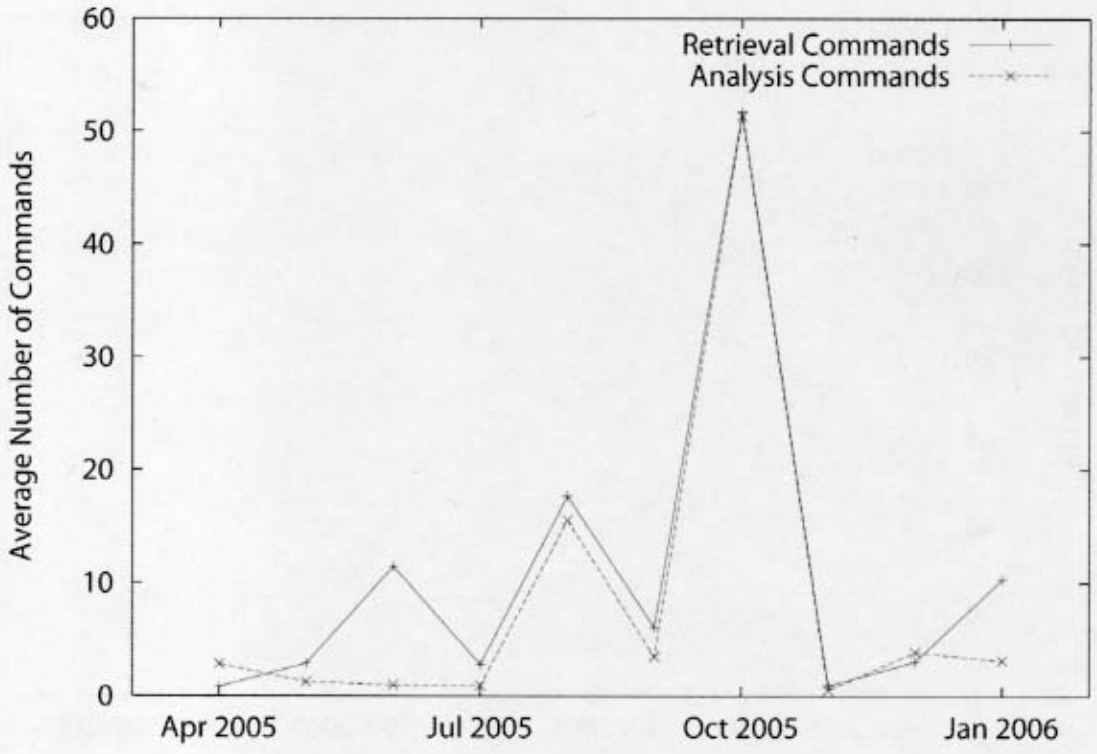


Fig. 12. Transition of average number of retrieval commands and average number of analysis commands in a single session.

by selecting the target data and the command in the manager window (Fig. 3). The coefficients window is displayed by selecting the command in the scattering chart window (Fig. 6).

The analysis operations are also executed through the menus. The user selects one or more retrieved data and then pushes the appropriate analysis command button on the menu. The analysis request is sent to the data server, which executes the operation to the selected data in the user's area. Figure 8 shows results of temporal averaging, and Fig. 9 shows results of diurnal averaging.

As stated above, the user interface is based on menu selection. Therefore, the user can conduct the analysis with simple operations.

4. Results of use

The data system for archiving and analysis is open to the CEOP community. Currently it has 27 users. It started service in April 2005 and had provided 517 user sessions by January 2006. The data system is based on Web service and the users do not explicitly disconnect when they stop accessing it. Accordingly, in this paper, we defined single session as starting from the login (user authentication) request and consisting of the leading login request and the following several requests from the same user other than the login request, such as retrieval requests, analysis requests, and so on. In other words, the session starts from the login request and ends with the request just before the next login request.

Figure 10 shows the number of user sessions of the data archiving and analysis system. Though there were 142 sessions in May 2005, there were 30 to 60 sessions per month after that and there has been no large change since. Figure 11 indicates the average time of a single session. Though the average time of a session fluctuates more than the number of sessions, the average times are becoming longer, roughly speaking. Figure 12 shows the average number of retrieval commands and of analysis commands issued in a single session. As the average numbers of retrieval and of analysis com-

mands increase, the average time of a session also increases. Session time increases as the numbers of retrieval and analysis commands increases. Analysis commands were issued as frequently as retrieval commands. In most cases, an analysis command was executed for retrieved data only once, though the executed analysis commands are not always the same.

The server program runs on a SUN Enterprise 6500. Though not the latest type of server, it has sufficient capacity to deal with current user requests.

5. Conclusion

This paper outlines the CEOP data archive system. We described the data to be archived and summarized the analysis functions. We also described the design and architecture of the data system in detail. Finally, we introduced the system's user interface. The preliminary data server is already operating and open to the CEOP community. We are now implementing more complicated analysis functions and improving the system. Though the data server has sufficient performance to process current user requests, we are trying to replace the server with the latest type to enable the system to cope with many more user requests. As well, as the centralized data archive system is based on a Web service, many techniques developed for Web services to improve performance are also applicable to the centralized data system.

The client program to access the data archiving and analysis server and its documentations are available at the CEOP data client homepage (http://monsoon.t.u-tokyo.ac.jp/ceop-dc/ceop-dc_top.htm).

References

Barclay, T., D.R. Slutz, and J. Gray, 2000: TerraServer: A Spatial Data Warehouse. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 307–318, Dallas, Texas, USA.

CEOP Data Client Homepage. http://monsoon.t.u-tokyo.ac.jp/ceop-dc/ceop-dc_top.htm.

A GUIDE TO THE CODE FORM FM 92-IX Ext. GRIB. http://www.wmo.ch/web/www/WDM/Guides/Guide-binary-2.html.

Koike, T., 2004: The Coordinated Enhanced Observing Period—an initial step for integrated global water cycle observation. *WMO Bulletin*, **53**(2), 115–121.

Marley, S., M. Moor, and B. Clark, 2003: Building Cost-Effective Remote Data Storage Capabilities for NASA's EOSDIS. *Proceedings of 20th IEEE / 11th NASA Goddard Conference on Mass Storage Systems and Technologies*, 28–39, San Diego, California, USA.

Nemoto, T. and M. Kitsuregawa, 2005: CEOP Data Server and Browse/Analysis Interface. *CEOP / IGWCO Joint Meeting Proceedings*, 87–90, Tokyo.

Nemoto, T., E. Ikoma, and M. Kitsuregawa, 2004: Design of data server for CEOP data. *Proceedings of the 2nd Asia Pacific Association of Hydrology and Water Resources Conference*, **2**, 558–565, Singapore.

NetCDF (network Common Data Format). http://www.unidata.ucar.edu/software/netcdf.

Stolte, E., C. Praun, G. Alonso, and T. Gross, 2003: Scientific Data Repositories—Designing for a Moving Target. *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 349–360, San Diego, California, USA.

Szalay, A.S., J. Gray, A.R. Thakar, P.Z. Kunszt, T. Malik, J. Raddick, C. Stoughton, and J. van den Berg, 2002: The SDSS SkyServer—Public Access to the Sloan Digital Sky Server Data. *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, 570–581, Madison, Wisconsin, USA.

Tamagawa, K., T. Nemoto, T. Koike, and M. Kitsuregawa, 2004: Introduction to the CEOP data integration system. *Proceedings CD-ROM (GAME CD-ROM Publication No. 11) of the 6th International Study Conference on GEWEX in Asia and GAME*, Kyoto, 01-07.