CEOP centralized data system and integrated analysis tools

Kenji Taniguchi[1], Toshihiro Nemoto[2], Eiji Ikoma[3], Masaki Yasukawa[4], Toshio Koike[5] and Masaru Kitsuregawa[6]

[1,5] Dept of Civil Eng, Univ. of Tokyo
[2,4,6] Institute of Industrial Science, Univ. of Tokyo
[3]Center for Spatial Information Science, Univ. of Tokyo
E-mail: taniguti@hydra.t.u-tokyo.ac.jp

Abstract

The amount of earth environmental data has increased explosively because of recent advances in observational techniques. On the CEOP (Coordinated Enhanced Observing Period) project, in order to improve our understanding of water and energy cycle, large amount of data is collected and archived. In this paper, we introduce the CEOP centralized data system and integrated analysis tool. The centralized system enables us to retrieve, browse, analyze and download CEOP data set, and it improves accessibility and usability of the CEOP data. Three dimensional data visualization system is developed with the fusion of various products and the cutting out of demanded arbitrary surface. Developing an easy-to-use interface for non-computational researchers. The CEOP data system is developed at Institute of Industrial Science, University of Tokyo. This paper also explains the design of the system.

1. Introduction

Water and energy cycle is important for peoples' life and the climate system. At the same time, there has been drastic progress in techniques for observing the earth environment, and as a result, the volume of earth observation data has increased. On the CEOP (Coordinated Enhanced Observing Period) project, large amount of data are being collected and archived in order to improve our understanding of water and energy cycle. CEOP data consist of three kinds of data, those are in-situ data, satellite data and model output data. The in-situ data are a temporal series of air temperature, pressure, humidity, precipitation and so on at 35 reference sites around the world. The satellite data are remotely sensed data from the operational satellites, such as TERRA, AQUA, TRMM, NOAA and so on. The model output data are generated by numerical weather prediction centers. These data have various dimensions, spatial and temporal resolutions, precision, formats, coordinate systems. The total amount of the data is almost 100TB per year. This paper explains the design and implementation of data server in addition to the browse and analysis interface which enable researchers to use our system easily.

It uses tape library system and disk arrays to store the data, however, the location of data is hidden from users. The users can retrieve data without considering data location. The browse and analysis interface is the client of the data server and it provides the users with menu based integrated access tools to the data server. The connection between the clients and the server is based on HTTP. The users can access all kinds of data through the same interface without taking account of data type. The users can view the retrieved data as graphic charts or bitmap images depend on their dimension directly from the server. Some analysis operations such as average, difference, correlation, and so on can be applied into one or more retrieved data on the server through the interface.

2. Data

2.1. CEOP data

Three kinds of data are planned to be archive on the CEOP project. They are in-situ data, model output data and satellite data.

The in-situ data are a 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 observation, subsurface observation and upper air observation. The surface observation consists of air temperature, pressure, humidity, precipitation, heat flux, radiation and so on at the ground level. The subsurface observation is composed of soil temperature, soil water content and soil heat flux for 2cm to 175cm depth. The upper air observation consists of air temperature, humidity, pressure and so on measured by radiosonde. The all values are not always observed at a reference site. The sorts of the observed values and observation frequency depend on the reference site. The total amount of in-situ data for two year and three months is almost 600MB.

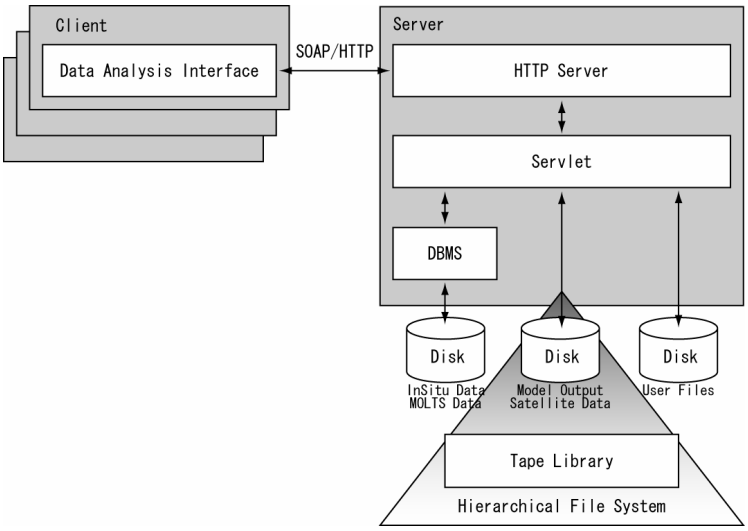

Figure 1

The model output data is the gridded values from global forecast model or assimilation system generated by 10 num erical weather forecast centers. Two types of model output, namely gridded data and site-specific time series at each of the reference site are planned to archive. The latter time series are designated as MOLTS (Model Output Location Time Series). The gridded data are three dimensional data and each cell has several prognostic variables such as air temperature, humidity, pressure and so on. The forecast length, assimilation intervals, the grid systems of the models and also the variables in models are different each other. The MOLTS data are one dimensional time series of variables at the reference site extracted from gridded data. The total amount of model output data is almost 20TB.
The satellite data are remotely sensed data from the sensors on the operational satellites such as DMSP SSM/I, TRMM TMI, TRMM PR, GMS S-VISSR, NOAA AVHRR, TERRA/AQUA MODIS, AQUA AMSR-E and so on. The satellite data are two dimensional data and each sensor has one or more channels. Though these data are geometrically corrected by the data supplier, their resolutions depend on the sensor and the channel. The observation time may vary day by day even if the same satellite. The total amount of satellite data is almost 200TB.

2.2. Data for integrated analysis tools

In addition to CEOP centralized data system, we have developed a prototype system of analysis tools for large volume data set.

For the analysis system, monthly average data for global precipitation from 1979 to 2000 and daily average data from 1997 to 2000, which are in the GPCP(Global Precipitation Climatology Project) data set. The resolution is 1 degree per pixel on every side. And, because OLR (Outgoing Long-wave Radiation) is useful to describe summer monsoons as an indication of convective activity in a tropical area, we use OLR data from 1975 to 2000 provided by NOAA (National Oceanic and Atmospheric Administrator). The resolution is 2.5 degrees on every side. Reanalysis data of NCEP/NCAR (National Center for Environmental Prediction / National Center for Atmospheric Research) for atmospheric data is archived and used. The resolution is also 2.5 degrees on every side and frequency is daily. We use sea surface temperature data provided by TMI (TRMM Microwave Imager) on TRMM (Tropical Rainfall Measuring Mission). This data has a 0.25 degree spatial resolution, and three day average temporal resolution from December 1997. Atmospheric Infrared Sounder (AIRS) onboard Aqua, NASA's satellite, is designed by NASA Jet Propulsion Laboratory. With other satellite sensors onboard Aqua, cloud effects can be reduced and vertical distribution of atmospheric temperature and water vapor and other atmospheric elements can be detected by AIRS. In this study, data of atmospheric temperature and water vapor are used. Spatial resolution is 0.25deg*0.25deg in horizontal direction, and 28 layers in vertical direction. Details are described in the web site of "AIRS data support".

3. System

3. 1. CEOP centralized data archiving system

- System architecture

The CEOP centralized data system is based on a client-server model. The architecture of the system is shown in Figure 1. The communication protocol between server and client is HTTP. The requests from clients are at first received at the HTTP server and then they are sent to the data manager. The data manager is a servlet program. It receives the requests from clients through the HTTP server and then it generates SQL commands for data search or executes analysis operations according to the user requests. One dimensional data such as in-situ data and MOLTS data are stored in DBMS, however, two and three dimensional data such as gridded model output data 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, the accessing a LOB in DBMS is slower than accessing a file. 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. Generally the values around the reference sites are necessary to compare the values from ground observations and that of model output or that of remotely sensed data. To store the small portions on the disks instead of hierarchical file system we can reduce the response time. The location of the data is hidden from the users. The users do not have to consider where the data are stored. The data server automatically migrates and retrieves the appropriate data from DBMS, disks or hierarchical file system as the user requests and sends them to the clients. The DBMS manages the one dimensional data and the metadata for all data. We use a commercial DBMS and JDBC for the connection between DBMS and the data manager.

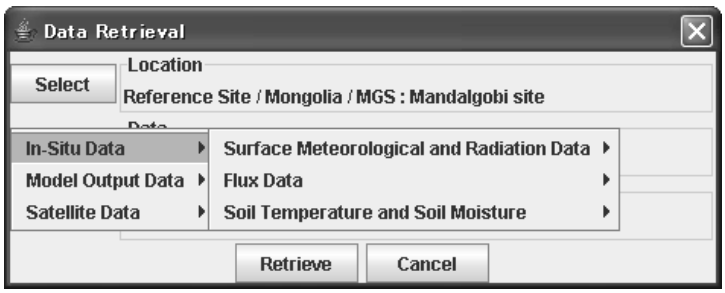

Figure 2

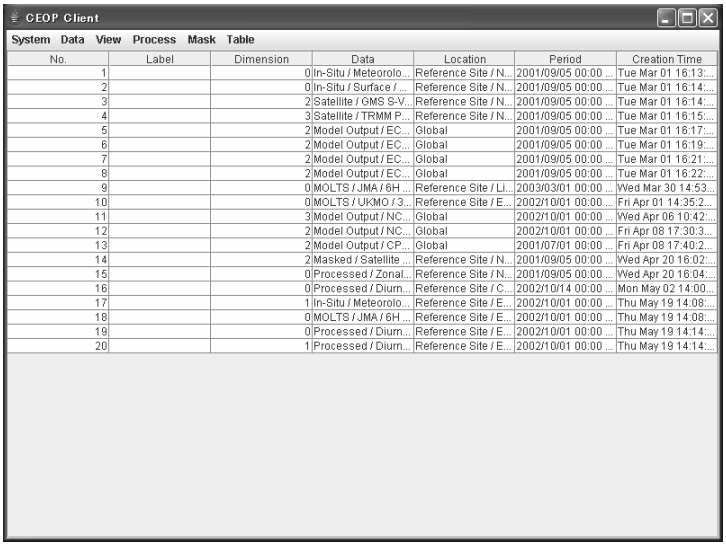

Figure 3

The graphical user interface (GUI) in the client system roughly consists of two parts, HTTP browser and data analysis interface. The communication between clients and servers are based on HTTP and the data analysis interface is written in JAVA. Accordingly, the client does not need any special hardware or software. Only HTTP browser and JAVA runtime environment are required. Since many kinds of current computers and operating systems support HTTP browser and JAVA runtime environment, the GUI for the data server works on many kinds of computers. The usage of the data analysis interface is described in the next section.

- Browse and analysis interface

The user accesses to the data server page through the HTTP browser at first and then authentication page is shown. After the user passes the password check, the user can access to the data listed in his/her work space. The requested data is specified by three items, location (reference site name), data name and temporal period. The available location and data name are listed in the menus. The user can select year and month in the menu as the temporal period or input the start date and time and the end of those (Figure 2). Clicking the button to retrieve, 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 portion into the user's area. The items in the data management window correspond to the result portions in the data server. The retrieved data can be displayed as a line chart (Figure 4) and a bitmap image (Figure 5). Users can specify any of the time, height, latitude and longitude as x and y axis of the graph or the image. When the result portion has more than three dimensions, the values in the axis except x or y axis can be changed by the slider bar on the bottom of the line graph window or the

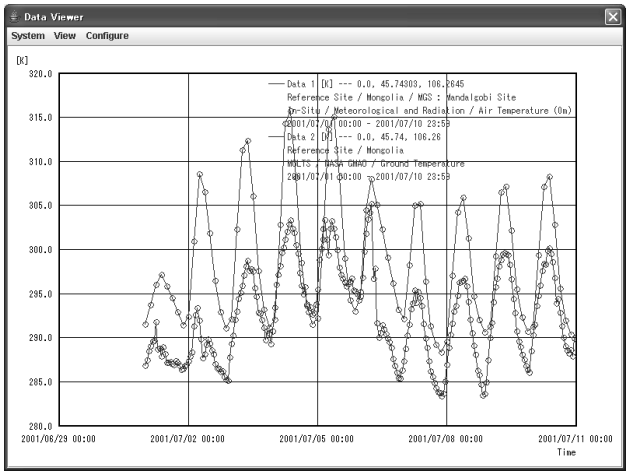

Figure 4

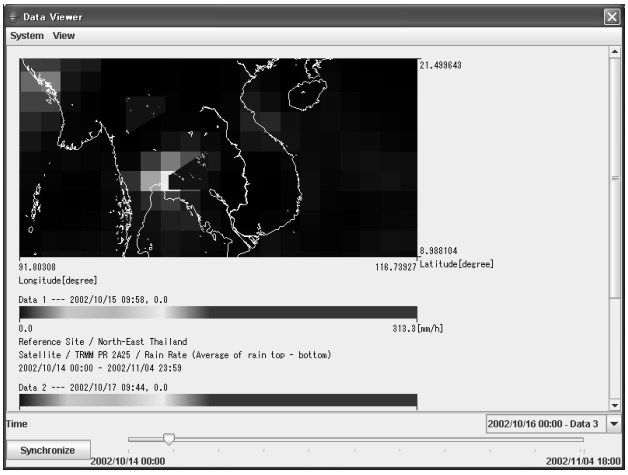

Figure 5

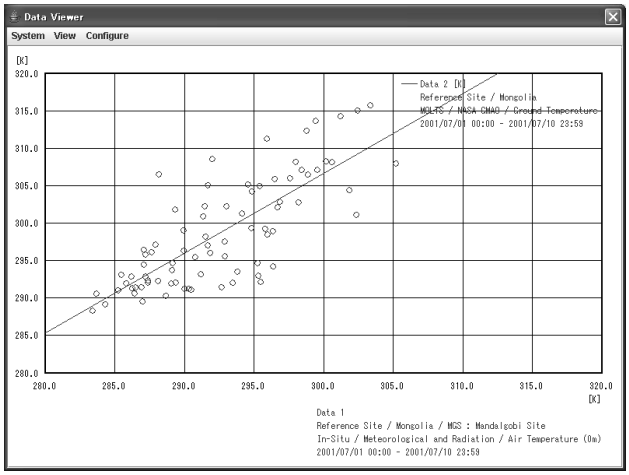

Figure 6

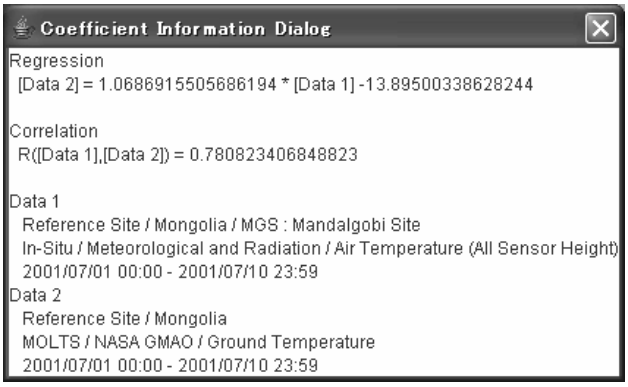

Figure 7

bitmap image window. For example, when the user specify longitude as x axis and latitude as y axis to generate the bitmap images of the temporal series of global sea surface temperature, he can change the time by the slider bar on the bottom of the window. A scatter diagram can also be drawn (Figure 6). In the diagram, the regression lines are also drawn. The regression coefficients and the correlation coefficients are shown in the coefficient information window (Figure 7).

The analysis operation is executed to the retrieved data. The user selects one or more retrieved data and then pushes the appropriate analysis command button. The analysis request is sent to the data server and the data manager executes the operation to the selected data in the user's area. In the current preliminary version, only simple analysis commands such as calculating average, maximum, minimum and so on are available (Figure 8). The data processed by the analysis operation are also stored into the user's area and they are listed in the data management window. They can be targets of analysis operations again. The user can apply the analysis operations onto the processed data again and again.

3.2. Integrated data integration tools

3.2.1. Correlation Coefficient Analysis Tools

Statistical analysis is used for extracting correlations between phenomena in meteorology. However, in the very large data sets that we targeted for this study, statistical analysis is quite difficult using general software. Moreover, the data we are targeting has not only simple correlations between two data, but also various other kinds of correlation such as spatial differences, temporal differences, and the difference of values; thus it requires tools which enable flexible handling and analysis and the ability to specify conditions. In this study, we introduce a tool which can compute correlation coefficient while specifying conditions such as time (term), space (area), value, temporal resolution and spatial resolution. With this tool, assuming that some natural phenomena may happen not at the same time but with some delay (time lag), a user can analyze the correlation between phenomena with different times (time lag correlation)

Using this method, users can find not only relationship between spatially separate points but also temporally separate points (such as a phenomenon happening a few days after the base phenomenon). We developed a GUI which can operate these tools via the Web. With this GUI, users can specify various conditions and visualize the results. It also allows multiple users to use it from remote locations.

When users log on our system, Top Page (Figure 8) will appear.

In the right frame, user can specify as follows:

- Choose 1 base data for correlation analysis
- Choose some (multiple) target data for correlation analysis
- Year, Month, Date
- Term
- Specify base area for base data
- Choose the window for displaying results (Same window of this page or other)

Then, the user's specified area on a world map will be displayed in the upper left frame for checking the area, and user will be asked the value of the threshold for visualization and the mode for processing. User can select whether results are immediately displayed or the system will notify user by e-mail after processing.

In this window, the results of different days of lag are lined up at horizontal direction, and the results of different value are lined up at vertical direction. In other words, one result shows correlation coefficients between spatial averaged time sequential and all points of the whole globe for a time lag period, year, month, date. Positive correlation points are shown as red points, and negative ones are shown as blue points, the darkness of color means it is strong and weak.

Such a display method enables users to understand change of correlation as time goes on and which data has a strong correlation with the base data by comparing the result at vertical direction.

If a user clicks interesting data on this window (Figure 9), a new window will appear and a higher resolution image is displayed, and the graph of area averaged time sequential data is displayed in the lower window. Clicking a point such as one with very high correlation on the global map of this window, more detailed information on each point and each time will be shown in new window.

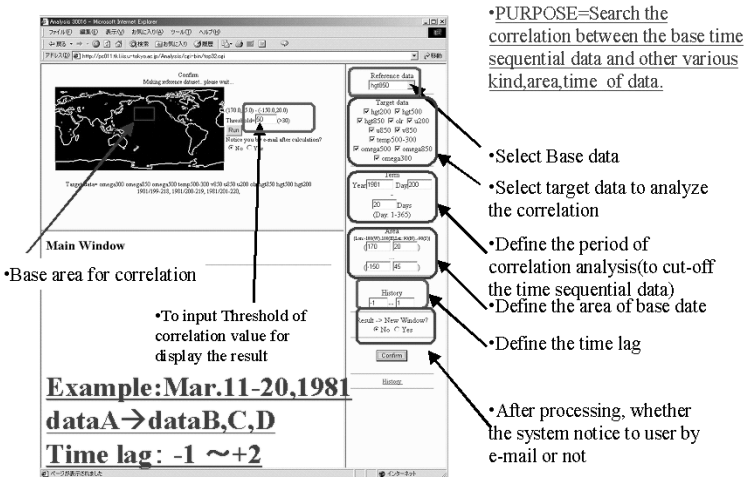

Figure 8

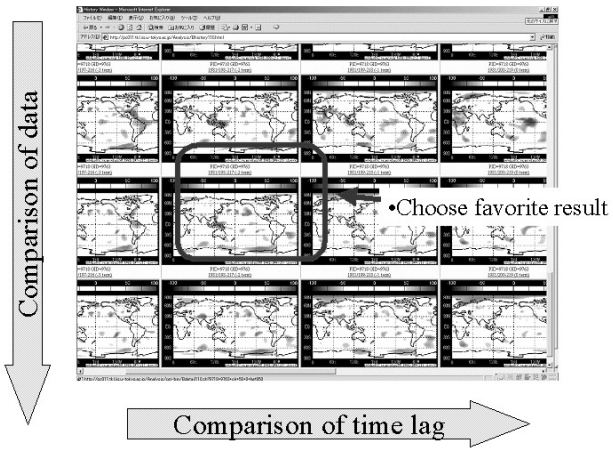

Figure 9

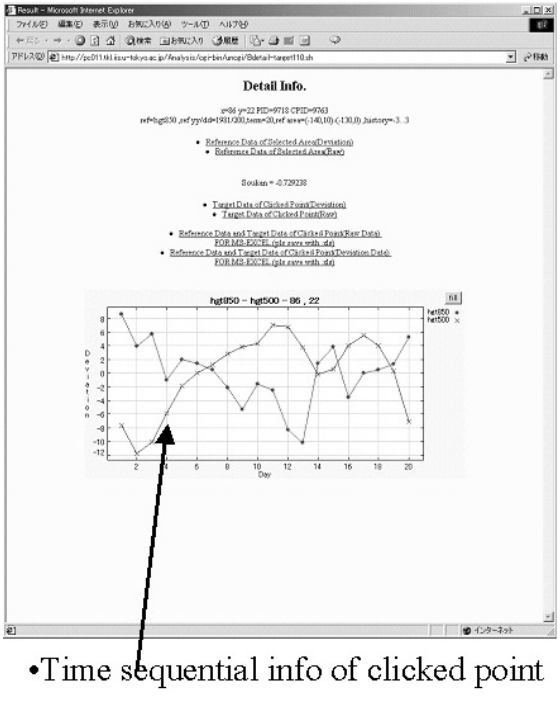

Figure 10

3.2.2 AIRS 3D Visualization system

-Basic design of AIRS visualization system

In this study, following two functions are required for the system;

-3D visualization

-Data subset and visualization for arbitrary curved surface

Two dimensional horizontal or vertical cross section is not sufficient method to utilize three dimensional data. For fully utilization of three dimensional data, virtual reality markup language (VRML) is used in this system. VRML does not require any commercial software and works on web browser, and it enables to see three dimensional data from arbitral view points.

At the same time, cross section is helpful to understand phenomena. However, software which have been used in meteorological research community has not allowed to make cross section in arbitral cross section. Sometimes, this limitation possibly leads misunderstandings of reality. Then, the second function is developed in this study.

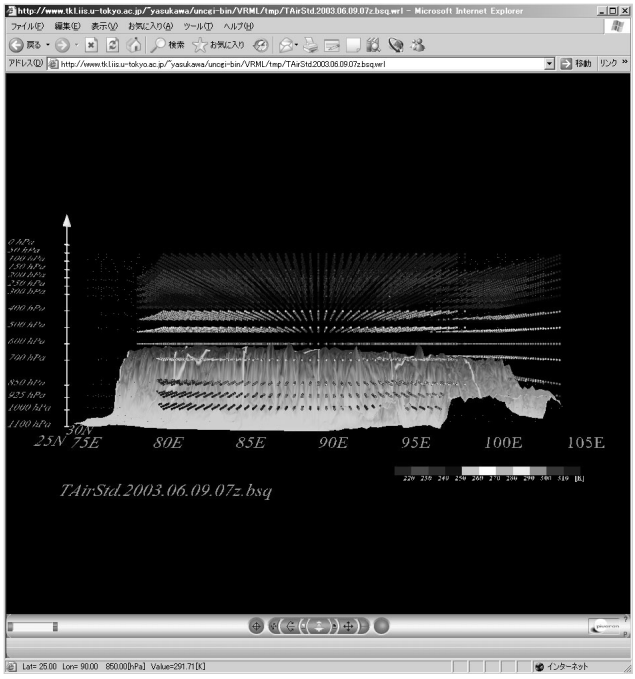

Figure 11

-VRML

In this system, three dimensional visualization of AIRS data is realized with web browser using VRML 2.0. By using VRML, users can view data from any direction and distance in virtual reality space. Overlaying geographical data, morphological feature and its effect on phenomena can be recognized at the same time (Figure 11). VRML also enables to plot multiple variables in the same space (Figure 12). Relationship among several elements can be recognized and it is useful for analysis.

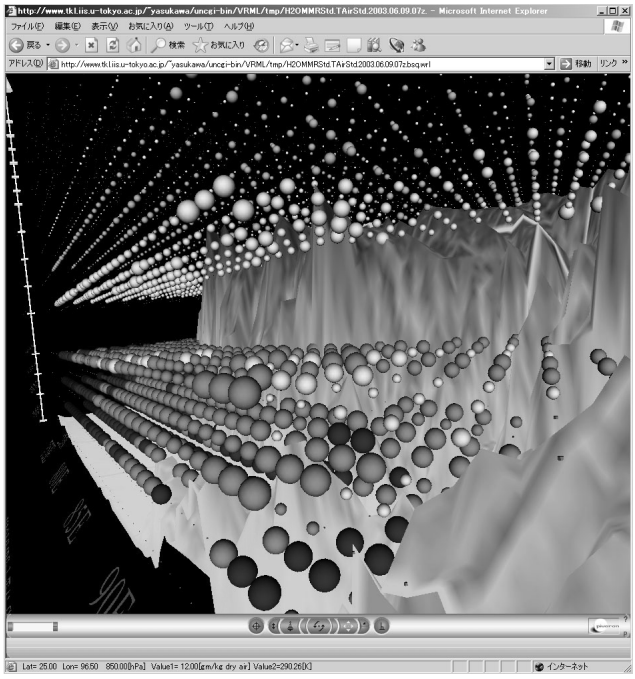

Figure 12

-Subset and plot on arbitral curved surface

In this system, users can select arbitral vertical section by inputting longitude and latitude on the user interface (Figure 13). First, data is retrieved from data server. Retrieved data is re-sampled by using selected curved cross section and visualized by GrADS. Images are made for selected period (Figure 14). These images are viewed as animation by choosing appropriate option. Plotted images can be overlay on VRML and viewed with geographical features. This system can also make horizontal 2D plot and overlay it in virtual reality space. These functions enable us to understand vertical distribution of a variable and horizontal 2D atmospheric condition in the same moment.

4. Summary

Progress in observation and monitoring technology in meteorology has enabled researchers to find new knowledge because of an increase in data volume. However, it has also brought some problems regarding methods and environments to use those huge volume data effectively and practically.

In this paper, centralized data system and integrated data using system for huge amount data developed by Institute of Industrial Science, University of Tokyo in CEOP activity were introduced.

CEOP centralized data system has managing functions for huge and various formatted data set. Users can retrieve, list and visualize data in their work space. At the same time, the data can be downloaded through the interface.

Moreover, the system enables us to process retrieved data with basic mathematics functions. The data server has been opened to the CEOP community tentatively.

At the same time, collaborating closely with the researchers who are using earth environmental data, analyzing tools for large volume data set are developed. Tools are based on web technologies for easy access and usage.

Though our current preliminary data server supports only simple operations, it may be helpful for studies related on water and energy cycle. We are now implementing more analysis functions and improving the system and also develop more flexible and effective tools with the feed back opinion from active users.

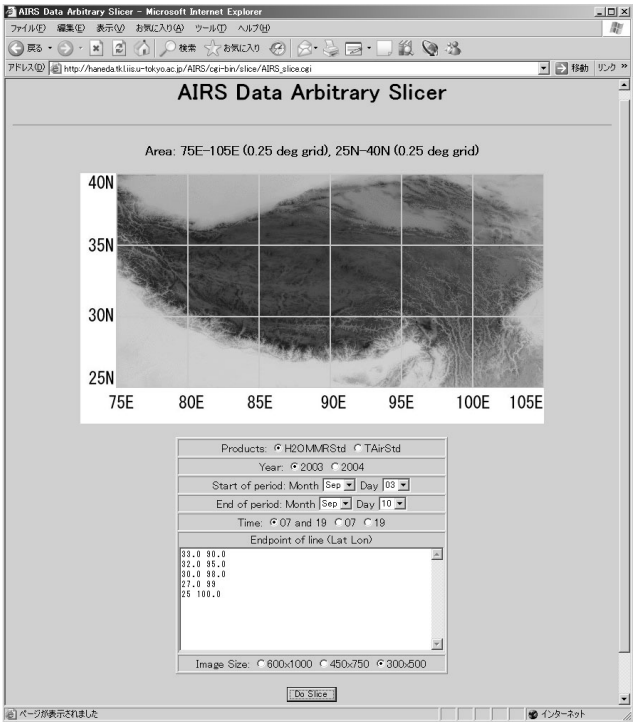

Figure 13

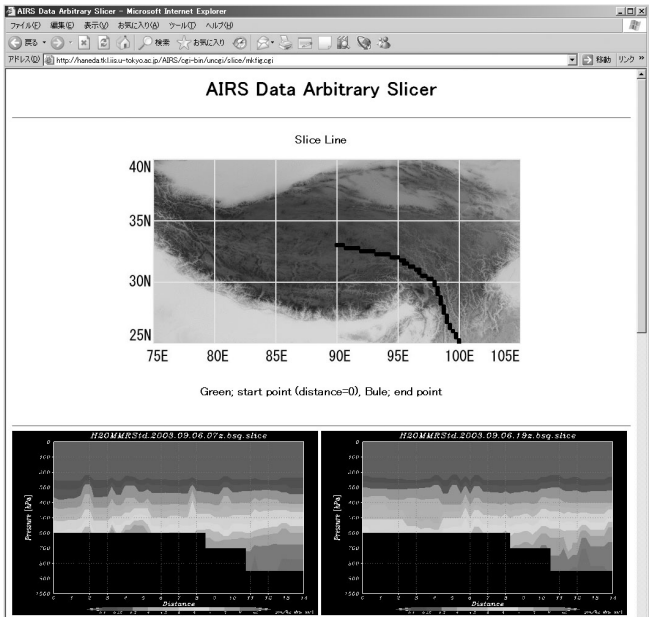

Figure 14

References

[1] Stolte E., Praun C., Alonso G. and Gross T.: Scientific Data Repositories – Designing for a Moving Target, *ACM SIGMOD 2003*.

[2] CEOP Home Page: http://monsoon.t.u-tokyo.ac.jp/ceop/.

[3] Nemoto T., Ikoma E. and Kitsuregawa M.: Design of data server for CEOP data, *Proceedings of the 2nd Asia Pacific Association of Hydrology and Water Resources Conference*, Vol.2, pp.558-565, 2004.

[4] Jet Propulsion Laboratory (NASA), "AIRS - Atmospheric Infrared Sounder -, "http://www-airs. jpl.nasa.gov/

[5] Jeanne Behnke, Alla Lake, "EOSDIS: Archive and Distribution Systems in the Year 2000, " Proceeding of 8th NASA Goddard Conference, pp.313-324, Mar. 2000.

[6] NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) / Distributed Active Archive Center (DAAC), "AIRS data support, " http://daac.gsfc.nasa.gov/atmodyn/airs/

[7] Kitsuregawa Lab., "Satellite Image Archive at Institute of Industrial Science, University of Tokyo", http://www.tkl.iis.u-tokyo.ac.jp/SIAIIS/

[8] Kitsuregawa Lab., "Data Visualization System for Earth Environmental Researchers, " http://www.tkl. iis.u-tokyo.ac.jp:8080/DV/

[9] U.S. Geological Survey (USGS), "GTOPO30 - Global Topographic Data -, " http://edcdaac.usgs. gov/gtopo30/gtopo30.asp

[10] Center for Ocean-Land-Atmosphere Studies (COLA), "GrADS Data Server (GDS), " http://grads.iges. org/grads/gds/gds.html

[11] Nemoto, T. and M. Kitsuregawa, CEOP Data Server and Browse/Analysis Interface, Proceedings of CEOP/IGWCO Joint Meeting, 2005

[12] Ikoma, E., K. Taniguchi, T. Koike and M. Kitsuregawa, Development of a Visual Data Mining Application for Earth Environmental Data, Proceedings of CEOP/IGWCO Joint Meeting, 2005

[13] Yasukawa. M., T. Nomoto, T. Koike and M. Kitsuregawa, Development of a Visualization System in Earth Observation Satellite Data with Three-dimensional Information such as AIRS Data, DBSJ Letters, vol.4 (1), 2005 (in Japanese)

[14] T. Koike, "The Coordinated Enhanced Observing Period – an initial step for integrated global water cycle observation", WMO Bulletin, vol.53, no.2, pp.115-121, April 2004.