Display Wall Empowered Visual Mining for CEOP Data Archive

Eiji IKOMA  
Center for Spatial Information Science, The University of Tokyo, Japan  
Masaru KITSUREGAWA  
Institute of Industrial Science, The University of Tokyo, Tokyo, Japan  
Kenji TANIGUCHI, and Toshio KOIKE  
Department of Civil Engineering, The University of Tokyo, Japan  
(Manuscript received 25 February 2006, in final form 19 December 2006)

Abstract

A large display space potentially improves data exploration, especially when there are several types of data sets to control. To better understand the feasibility of a large display space, we implemented a display wall assisted visual mining system. A 5 m × 2.5 m display allows users to simultaneously visualize various types of data. The system also has a very high resolution of 15 XGA displays (each XGA display has 1024 pixel × 768 pixel resolution). With these attributes, the user can easily examine and visualize a dense data set in detail. To our knowledge, few systems of this type that are easy to use have been realized. A number of systems are now functional (e.g., the NASA AMES Hyperwall, http://www.nas.nasa.gov/Groups/VisTech/hyperwall/). However, these systems require much knowledge and power of the user. Most users of such systems have to use some special application, which can be difficult for earth environmental researchers. Although the proposed system has had limited testing by some researchers, we believe that our experience would be helpful for users who are trying to develop similar systems. The proposed system is more efficient as it allows for simultaneous multi-dimensional visualization of complex data products, and various mining functions, such as two-dimensional correlation, are also prepared for the Coordinated Enhanced Observing Period (CEOP) data archive. By using these functions, various advanced analyses that are not easily utilized with conventional mining tools are attained. Moreover, several scenarios using the powerful capability of this tool are also presented. Finally, the process of data analysis in the proposed system becomes much more efficient through the use of a large-scale high-resolution display wall.

1. Introduction

There has recently been extensive progress in techniques for observing the Earth’s environment. As a result, the volume of remote sensing data, such as satellite and point data acquired from various types of land observation instruments has increased. Large amounts of useful and detailed data, which were previously difficult to obtain, are now available. This contributes to progress in various fields of research. However, because of the sheer volume of new data, researchers have not been able to process much of this data using traditional analytical techniques. Those not familiar with or not having access to computers have been unable to
use the new data. Much of the important information collected over the past few years is simply being stored and left untouched at various data archive centers. Moreover, in many cases, even with some specialized tools there is no display capability with sufficient visualization power to exam the data. Detailed examinations have, therefore, not been performed even if the amount of data acquired or generated is huge.

In the present study, in collaboration with researchers who are using these enormous amounts of earth environmental data in practical ways, we have developed a visual mining system that targets the real data needed to support on-going analyses. We also developed a Web-based user interface system that enables researchers who are not familiar with computers to easily use the proposed system. Finally, we provided a practical research example using the tools developed in the present study.

2. Background

2.1 Increase in available earth environmental data and archival methods

The success of data collection using satellite remote sensing has lead to an increase in the number of sensing satellites. It has also resulted in the development of various new types of sensors. Thus the volume of earth environmental data acquired in this manner has increased. Moreover, in the fields of meteorology and climate change research, improvements in computer technology have enabled the analysis of climate forecasts and the creation of simulations with higher spatial and time resolutions. However, while the total amount of data available is likely to increase each year, the amount of data that researchers can handle each year is limited. Under such circumstances, a great deal of data is going to be left in archival format and not used for research. The Japan Aerospace Exploration Agency (JAXA), the National Aeronautics and Space Administration (NASA), the European Space Agency (ESA), and the climate forecast centers in Asia, Europe, and the Americas have collaborated to establish the Coordinated Enhanced Observing Period (CEOP). CEOP Phase 1 includes four so-called Enhanced Observing Periods (EOP1–EOP4). During EOP3 (October 2002–September 2003) and EOP4 (October 2003–December 2004) substantial amounts of data, including satellite (remotely sensed) data, predicted data from each of the participating numerical prediction modeling centers, earth observed data collected by space agencies, and various other types of data generated by various projects and organizations were collected.

The amount of satellite data collected during the CEOP period was almost 90 terabytes a year, with numerical prediction and objective analyzed data comprising approximately 55 terabytes, with more observed data continuing to be collected. However, research into how best to analyze this huge amount of data, including the extraction of important information and the acquisition of new knowledge, has not been adequately progressed.

2.2 Visualization on a large display

Advances in technological fields are making it possible to produce higher-resolution visualization equipment for application in various research fields. Such advances are expected to continue into the future.

Compared to visualization using a display connected to a general desktop PC or the use of a liquid crystal display monitor of a notebook PC, visualization on a large display has two advantages.

The first is the ability to visualize all of the information at one time, something that was not previously possible. In the past, information had to be viewed a little at a time because of the screen size. Currently, several small objects can be simultaneously displayed on a large display. Depending on the type of data, information that was previously displayed on separate screens might gain new correlativity and relevance that might not have been apparent until being visualized all at once.

For example, Fig. 1 shows sea surface temperature (SST) data displayed for tens of years on the proposed display wall. The data for one year is put in order in lines, and the same weeks of each year are displayed together. Thus, by displaying the data as a matrix, the change and the yearly variation over a long period of time, which was difficult to recognize previously, can be grasped easily in the conventional format. Moreover, being able to move around in front of the display wall is a big aid for researchers to better understand the data.

Another advantage is the ability to display
an object with high resolution, which can not be achieved on a conventional display. Information was checked, after reducing resolution or thinning out information because of the limit of the visualization equipment; even if the data originally had the information at high resolution. However, by visualizing the result at or near the original resolution, the discovery of previously overlooked information can be expected. Figure 2 shows an example of the display of a meteorological satellite picture having a horizontal resolution that exceeds 5,000 pixels on the display wall. Conventionally, researchers viewed only an area of interest, or they reduced the resolution and viewed a larger area. However, by displaying all of the data at once, the relevance of surrounding areas is more easily realized.

Thus, the proposed system is thought to be important for future research because the large high-resolution display will help generate new knowledge that previously was not noticeable.

3. Correlation analysis tools

Correlation analysis is an effective method for understanding natural phenomena, as described above. In this section, research concerning a Pacific North America (PNA) teleconnection is described as an example of utilizing correlation analysis. A data mining tool for Web-based development using a large display is also described.

3.1 Correlation of natural phenomena

In meteorology, various methods have been used to abstract correlations between phenomena by analyzing statistics. The most used method is to find a correlation coefficient. For example, Wallace and Gutzler (1981) checked the distribution of correlation coefficients based on the fluctuation of the height of 500 hPa at the lattice point of (45°N, 165°E). As a result, it became clear that a positive correlation area was located in the North Pacific Ocean, from the east coast to the central area of North America, with a negative correlation area located from the west coast to the central area of North America.

Figure 3 shows that in cases where the air pressure at the base lattice point becomes high, the air pressure at the positive (negative) correlation area becomes high (low). Since such distributions of correlation coefficients are widespread from the Pacific Ocean to the North
American coast, this region has been dubbed the Pacific North America (PNA) Teleconnection Pattern. Currently, this phenomenon is understood as a stationary Rossby wave that is dynamically energized by a heat source in the equatorial zone, so the detection using the analysis of correlation is explained conceptually. In this way, the traditional approach is to discover a phenomenon and try to formulate a theory by which it can be explained. A more recent approach would be to use models and theories to actively search for correlations between phenomena.

3.2 Development of web-based correlation analysis tools

As mentioned in Section 3.1, statistical analysis is the method of choice for extracting correlations between phenomena in meteorology. However, in the very large data sets targeted by the present study, statistical analysis is quite difficult using the usually available software. Moreover, the targeted data has not only simple correlations between data, but also has various other types of correlations such as spatial differences, temporal differences, and differences in values. Thus, tools that enable flexible handling and analysis and the ability to specify conditions are required. In the present study, we developed tools that can analyze correlations while specifying conditions such as time (term), space (area), value, temporal resolution, and spatial resolution, as well as coordinating with an immense data archive system. With these tools, assuming that some natural phenomena may not occur at the same time as the base phenomenon, but rather may occur with some delay (time lag), a user can analyze correlations between phenomena at different times (time lag correlation) (see Fig. 4).

Example:


1. Calculate the spatial average of the base area and set time sequential data as the correlation to the base area.

Fig. 2. High-resolution photograph on a display wall.
of the other data (target data) B, C, and D. If these data have 1-degree of spatial resolution, each data has $360 \times 180$ points.

3. Calculate the correlation between the base data and all target data, and generate the correlation coefficient of each point.


5. Calculate the correlation between the base data and the target data with time lag.

6. Such above-mentioned calculation results are visualized on a large visual display in the display style the user selects.

7. By clicking an interesting result, more detail information such as time series graphs etc. can be displayed.

Using this method, users can find not only correlations between spatially separate points but also those between temporally separate points (such as a phenomenon happening a few days after the base phenomenon). We developed a Graphical User Interface (GUI) that can be used to operate such tools via the Internet. This system is developed as the front stage of a system using a large display, which can be operated with an interface that is similar to that of normal Web browser through the Internet. Using this GUI, users can specify various conditions and easily visualize the results. As well, multiple users can use the GUI from remote locations.

4. Implementation of visual mining systems on the CEOP data archive

Recent advances in information technology have resulted in computers that quickly process huge amounts of data. However, as mentioned
earlier, when users attempt to clarify natural phenomena using earth environmental data, one major problem is that it is not easy to use these computers for higher-order computing or analysis. Indeed, most users who want to analyze data are not sufficiently skilled on computers to use anything but low-level, off-the-shelf software on a standard personal com-
puter. This means that even though there is a huge amount of data available, spatial and temporal resolution must be drastically decreased to use this data. This results in the loss of information or the misinterpretation of data trends. In the present study, collaborating closely with researchers investigating the Earth’s environment, we developed tools that can analyze huge amounts of Earth environmental data using very large data archive systems and a display wall. A system that has a web-based interface that enables users to access the system in their own research environments was also developed. By effectively visualizing results, the proposed system supports the visual cognitive capability and reasoning ability of humans, making it possible to more easily perform effective mining processing. In this section, we describe the type of data that can be used with the proposed system, newly developed analysis tools, the implementation/application of the display wall, its Web-based user interface, the system structure, and the contribution of the proposed system to the CEOP.

As well, research into the phenomenon of the Indian monsoon season is presented in the Appendix as an actual example of research using the proposed system.

4.1 Target phenomenon

The summer Asian monsoon has tremendous impact on a region of the world that has a large, concentrated, and rapidly expanding population. However, most of the physical processes of seasonal progressions or yearly variations in the monsoons in that region have not been fully explained.

Summer Indian monsoons have been studied for some time, and it is understood that a strong west wind that blows across the Indian subcontinent from the Indian Ocean and the
Arabian Sea brings rain by carrying moisture vapor to the Indian subcontinent. However, the process that builds the west wind is not well understood. One reason for this is that the temporal and spatial resolution of available data has not been high, restricting the ability to adequately trace the daily variations. Recent progress in sensing techniques and information technology has improved the quality and quantity of available data. In the present study, as an example of utilizing these huge amounts of useful data, we work on analyzing the process of the Indian Monsoon, focusing on its time of onset.

4.2 Using data
For analyzing the distribution of precipitation in the summer monsoon period as well as seasonal changes, we use monthly average data for global precipitation from 1979 to 2000, and daily average data from 1997 to 2000, which are included in the Global Precipitation Climatology Project (GPCP) dataset. The resolution is 1 degree per pixel on each side. Because Outgoing Long-wave Radiation (OLR) is useful for describing summer monsoons as an indication of convective activity in a tropical area, we use OLR data from 1975 to 2000 as provided by the National Oceanic and Atmospheric Administration (NOAA). The resolution is 2.5 degrees on each side. We use the sea surface temperature data provided by TRMM Microwave Imager (TMI) on the Tropical Rainfall Measuring Mission (TRMM). This data has a 0.25-degree spatial resolution, and a three-day average temporal resolution from December 1997. The total amount of data is approximately 900 GB.

4.3 System structure
The system structure is shown in Fig. 5. The condition and processing order specified through the web interface by the user are sent to data mining processing and data visualization processing through a web server. Data mining processing generates Structured Query Language (SQL) based on the conditions sent by the user and sends the request for data to a database in the data archive system. The acquired data is used for correlation analysis calculations in the data mining process based on these conditions. The results are then subjected to data visualization processing. Here, the data is processed for visualization based on the conditions sent from the web server, and effectively visualized results are shown on the web interface via the web server, where the user can view the results using a web interface. While

<table>
<thead>
<tr>
<th>WebServer, Data Archive Server</th>
<th>Sun Enterprise 6500 (UltraSPARC-336 MHz × 8.4 GB Memory)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining Server</td>
<td>Dell Poweredge1600c (Redhat Linux9, Xeon-3.2 GHz × 2.4 GB Memory)</td>
</tr>
<tr>
<td>Data Visualization Server</td>
<td>SGI ONYX4 Ultimate Vision (IRIX6.5.800 MHz-R16000 × 10, 4 GB Memory)</td>
</tr>
<tr>
<td>Data Archiver</td>
<td>Sun StorEdge A5000 (250 GB), STK WolfCreek 9960 + RedWood SD-3 System (0.5 PB)</td>
</tr>
<tr>
<td>Visualization Display</td>
<td>MITSUBISHI LVP-FD10 (50 inch × 15)</td>
</tr>
<tr>
<td>RDBMS</td>
<td>PostgreSQL 6.5.4</td>
</tr>
<tr>
<td>other</td>
<td>apache httpd, csh, awk, sed, c, java1.0, vrml2.0, etc.</td>
</tr>
</tbody>
</table>
the user can effectively check the visualized results, their analysis can be continued using their own visual cognitive and reasoning capabilities, as well as their own empirical domain knowledge.

The hardware and software used in this system are listed in Table 1.

4.4 Web-based correlation analysis system

When users login to the proposed system, a Top Page (Fig. 6) appears.

In the right frame, the user can specify the following attributes:

- WebServer, Data Archive Server
- Choose 1 base datum for correlation analysis
- Choose some (multiple) target data for correlation analysis
- Year, Month, Date
- Term
- Specify base area for base data
- Choose a window to display the results (Same window as this page or another)

The user-specified area will then be displayed on a world map in the upper-left frame to allow the area to be checked. At the same time, the user will be asked the value of the threshold for visualization and the mode for processing (whether the system will start the process and display results immediately or will notify the user by e-mail of the results after processing). After responding and pushing the start button, the process starts and the results (Fig. 7) are displayed. In this window, the results of different lag days are arranged horizontally, whereas the results of different values are arranged vertically. That is, the outcome dataset is the result of the correlation coefficients calculated between the time sequential area average data of the base area of the database, and all points in the world for one target area for one time lag period, year, month, and date. Positive correlation points are shown as red points, and negative ones are shown as blue. The shades of the colors signify whether the correlation is strong or weak. Using this method, the user can recognize changes in the correlations with the passage of time by comparing the horizontal results, which signify the degree to which the time lag affects the results, and by comparing the vertical results, which identify data that have strong correlations with the base data.

If the user clicks on specific data in this window (Fig. 8), the window shown in Fig. 9 is displayed.

In this window, the chosen correlation analysis results are displayed at a higher resolution, and the graph of the area average time sequential data of the base data is displayed in the lower window. If the user clicks on a point on the global map of this window, such as a
point having a very high correlation, Fig. 9 will be displayed to confirm more detailed information regarding the point and each specific time.

4.5 Implementation on the display wall

We assume that the proposed system will be used by researchers who are not familiar with high-performance computers. Thus, we use a web browser for visualization because it is easy to use and is found on most personal computers. However, as mentioned earlier, since visualization on a large high-resolution display is a very powerful tool for research, we are also developing a similar expanded system on a large display (Display Wall, Fig. 10). Since the display wall can allow for much higher resolution than a standard size computer screen, we can demonstrate how users can examine data in much more detail, or can simply examine much more data at one time. The proposed system is still under development with the help of researchers who have expertise in Earth environmental science. However, even in these early stages, these researchers were able to identify various new phenomena using the proposed system.

The simultaneous correlation and lag correlation coefficients between the air temperature over the western part of the Tibetan Plateau (TTW) and other physical variables were calculated and compared for each year from 1981 to 2000. Positive correlation was found between the TTW and the upward flow over the Arabian Sea. However, a negative correlation was found between the TTW data and the geopotential height at 850 hPa around the southern periphery of the Himalayas. The physical mechanism and its role in the Asian monsoon system are now under investigation.

This discovery was attained using the proposed system, which provided high-speed processing and quick visualization of a number of results. The large display wall enabled comparison of several visualized results. It is difficult
and time-consuming to derive these results in a general research environment; this new tool actively supports such research activities at a much higher and noire efficient level than with the usual equipment.

As mentioned in Section 2.1, the proposed system is being developed in collaboration with the CEOP Project and uses data collected by CEOP members (Section 4.2). Indeed, the development of the proposed system was initiated at the request of the CEOP participants. Currently, a number of researchers from the CEOP community are using the proposed system and have already realized numerous benefits of the system for Earth environmental research. In the appendix, we briefly describe such an example.

5. Conclusion

Progress in sensing technology in the field of climatology and meteorology, which uses earth environmental data, has enabled researchers to uncover new information because of an increase in data volume. The system has also highlighted a number of problems regarding the methods used to effectively examine such large quantities of data and the environments in which they can be effectively examined. In the present study, collaborating closely with researchers using Earth environmental data, we
• Time sequential info of clicked point

Fig. 9. Visualized results of a detail analysis on a display wall.

Fig. 10. Visual mining system on a display wall.
used real data with the proposed very large-scale data archiving system and developed analysis tools with data mining techniques. We also developed a web-based interface to support specialized analyses by building special environments for researchers. As a result, research concerning the summer Indian monsoons revealed two previously unknown seasonal progress patterns required for the onset process. In the future, we intend to develop analysis tools that will enable additional high dimensional methods. We will also develop more flexible and effective tools using feedback from active users. Moreover, as practical research using the proposed system is progressing, we will also learn from its usage logs.

Acknowledgements

This research is partially supported by ‘Development of Data Integration and Information Fusion Infrastructure for Earth Observation.’ Funded by the Ministry of Education, Culture, Sports, Science and Technology.

Appendix

Applications for practical research—analysis of the Indian monsoon

In this appendix, research concerning the Indian monsoon is carried out as part of the Earth Water Cycle Informatics Project.

1. Background

To plan stable and effective water resources management, improving the accuracy of longitudinal climate change forecast in medium- and short-term weather forecasts is required. The summer Asian monsoon has an immense impact, but most of the physical processes of seasonal progressions and yearly variations are poorly understood. Therefore, because its processes are somewhat better known, we start our example analysis using the Indian monsoon. We focus on the start of the Indian monsoon season using the proposed tools and various types of useful data that have become available as a result of remote sensing techniques.

2. Outline of the research

In the present study, we analyze the example in detail using approximately 22 years of multivariate daily average data inputed into the proposed system using the above-described tools. Details are discussed in Taniguchi et al. (2004).

As a result of our analysis of the west wind, which brings the rain in the summer Indian monsoon season, it has become clear that there are two factors involved in the start of the monsoon season. One is the change in the atmospheric field associated with a cyclone. The other is the generation of a field with a gradual building of the westerly winds, without any special event. Table 2 shows the results of our analysis of the seasonal changes until the constitution of the west wind from 1981 to 2002, and whether the wind results from a cyclone. The start time of the monsoon is defined as the day of the constitution of the west wind, which is also the day when the average wind speed of the area is at lat. 10°N–lat. 15°N, lon. 60°E–lon. 70°E and 850 hPa. It also refers to the plot data of the wind system.

There are nine examples in this 22-year period when the start of the west wind is caused by cyclone activity. Cyclonic circulation causes a rapid change in the atmospheric field, but the time and place of onset differ each year. On
the other hand, where the cause is not cyclonic, there are a number of common characteristic points in the seasonal progression in 2001. Briefly, the sequence is as follows, a northwest wind over the west coast of Middle East Asia and the Indian subcontinent is generated by the heat of the Indian subcontinent and the Iran-Pakistan area, this event carries warm air over land and is then heated over the Arabian Sea. The heat of the Arabian Peninsula increases. The southwest wind, which blows from Eastern Africa to the Arabian Sea due to the temperature inclination between the Arabian Sea and Peninsula, then forms. Finally, the temperature of the Arabian Sea falls again. By increasing the temperature inclination between the warm area (Arabian Peninsula area and Iran-Pakistan area) and the cold area (Arabian Sea), a strong west wind is generated on the Arabian Sea. In other words, for monsoons to be generated, these two temperature inclinations between the Arabian Peninsula and Arabian Sea, and between Middle-East Asia and the Arabian Sea, are necessary to build up the west wind over India.

Figure 11 shows the onset day for monsoons caused by the two factors described above.

The temperature inclination of the monsoon onset when there is cyclonic action is smaller than when there is no cyclonic action. That is, under conditions in which the temperature inclination becomes large enough, a west wind has already been generated. Whether or not a cyclone arises has an influence on the two patterns of west wind generation. Consequently, we learned that if we can predict the rise of cyclones when the temperature inclination is not above the required threshold, we may be able to predict the initiation of the west wind, which in turn signifies the beginning of an Indian monsoon.

References

ACESgrid Hyperwall (http://acesgrid.org/hyperwall_main.html).
Chelliah, M. and P. Arkin, 1992: Large-scale vari-

CEOP Homepage (http://www.ceop.net).

NASA Ames Hyperwall (http://www.nasa.gov/Groups/VisTech/hyperwall/).

National Centers for Environmental Prediction (http://ncep.noaa.gov/).


