Fast Spatio-Temporal Data Mining of Large Geophysical Datasets

P. Stolorz
Jet Propulsion Laboratory
California Institute of Technology

H. Nakamura
Dept. of Earth and Planetary Physics
University of Tokyo

E. Mesrobian, R. R. Muntz, E. C. Shek
J. R. Santos, J. Yi, K. Ng, S.-Y. Chien
Computer Science Dept., UCLA

C. R. Mechoso
J. D. Farrara
Atmospheric Sciences Dept., UCLA

Abstract

The important scientific challenge of understanding global climate change is one that clearly requires the application of knowledge discovery and datamining techniques on a massive scale. Advances in parallel supercomputing technology, enabling high-resolution modeling, as well as in sensor technology, allowing data capture on an unprecedented scale, conspire to overwhelm present-day analysis approaches. We present here early experiences with a prototype exploratory data analysis environment, CONQUEST, designed to provide content-based access to such massive scientific datasets. CONQUEST (CONtent-based Querying in Space and Time) employs a combination of workstations and massively parallel processors (MPP's) to mine geophysical datasets possessing a prominent temporal component. It is designed to enable complex multi-modal interactive querying and knowledge discovery, while simultaneously coping with the extraordinary computational demands posed by the scope of the datasets involved. After outlining a working prototype, we concentrate here on the description of several associated feature extraction algorithms implemented on MPP platforms, together with some typical results.

Introduction

Understanding the long-term behavior of the earth's atmospheres and oceans is one of a number of ambitious scientific and technological challenges which have been classified as "Grand Challenge" problems. These problems share in common the need for the application of enormous computational resources. Substantial progress has of course already been made on global climate analysis over the years, due on the one hand to the development of ever more sophisticated sensors and data-collection devices, and on the other to the implementation and analysis of large-scale models on supercomputers. Gigabytes of data can now be generated with relative ease for a variety of important geophysical variables over long time scales. However, this very success has created a new problem: how do we store, manage, access and interpret the vast quantities of information now at our disposal?

The issue of data management and analysis is in itself a Grand Challenge which must be addressed if the production of real and synthetic data on a large scale is to prove truly useful. We present here early experiences with the design, implementation and application of CONQUEST (CONtent-based QUerying in Space and Time), a distributed parallel querying and analysis environment developed to address this challenge in a geoscientific setting. The basic idea of CONQUEST is to supply a knowledge discovery environment which allows geophysical scientists to 1) easily formulate queries of interest, especially the generation of content-based indices dependant on both "specified" and "emergent" spatio-temporal patterns, 2) execute these queries rapidly on massive datasets, 3) visualize the results, and 4) rapidly and interactively infer and explore new hypotheses by supporting complex compound queries (in general, these queries depend not only on the different datasets themselves, but also on content-based indices supplied by the answers to previous queries). CONQUEST has been built under the auspices of NASA's High Performance Computing and Communications (HPCC) program. Although it is geared initially to the analysis and exploration of atmospheric and oceanographic datasets, we expect that many of its features will be of use to knowledge discovery systems in several other disciplines.

Content-based access to image databases is a rapidly developing field with applications to a number of different scientific, engineering and financial problems. A sampling may be found in volumes such as (Knuth & Wagner 1992, Chang & Hsu 1992). One example is the QUBIC project (Niblack et al. 1993) illustrating the state-of-the-art in image retrieval by content. Examples of work in the area of geoscience databases include JARTool (Fayyad et al. 1994), VIMSYS (Gupta, Weymouth & Jain 1991) and Sequoia 2000 (Guptill & Stonebraker 1992). Many of these efforts are directed at datasets which contain relatively static high-resolution spatial patterns, such as high-resolution Landsat imagery, and Synthetic Aperture

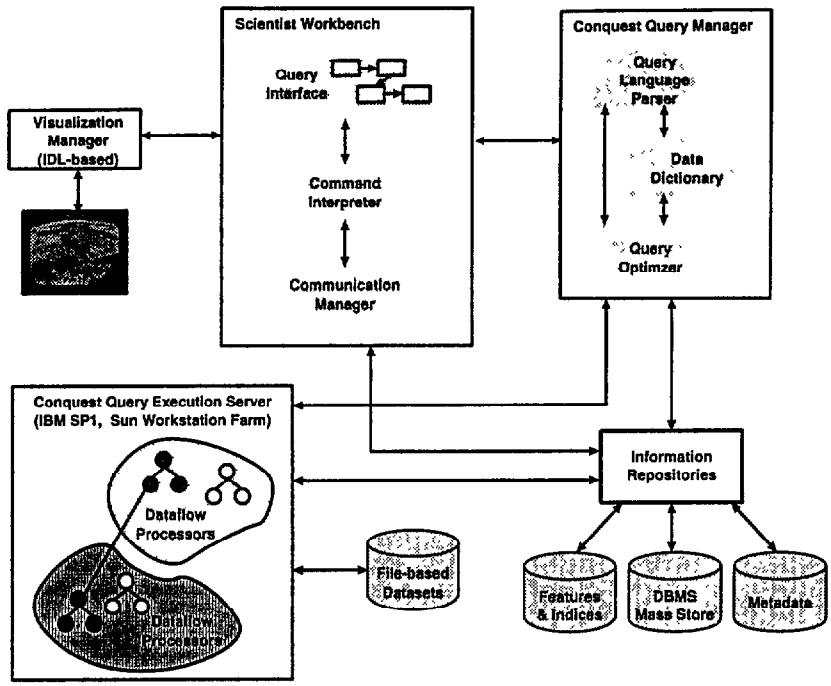

Figure 1: System Architecture

Radar imagery of the earth's surface and of other planets. CONQUEST shares a great deal in common with these systems. Its distinguishing features are, 1) the fact that it is designed to address datasets with prominent temporal components in addition to significant high-resolution spatial information, and 2) that it is designed from the beginning to take maximum advantage of parallel and distributed processing power.

The remainder of this paper is laid out as follows. We first introduce our underlying system architecture, followed by an introduction to the datasets that have been used as a testbed for the system. We then focus on the extraction of two important spatio-temporal patterns from these datasets, namely cyclone tracks and blocking features, outline an algorithm to perform hierarchical cluster analysis, and describe the efficient implementation of these procedures on massively parallel supercomputers. These patterns exemplify the types of high-level queries with which CONQUEST will be populated.

System Architecture

The system architecture is outlined in Figure 1. It consists of the the following 5 basic components:

- Scientist Workbench
- Query Manager (parser and optimizer)
- Visualization Manager
- Query execution engine
- Information Repository

The scientific workbench consists of a graphical user interface enabling the formulation of queries in terms of imagery presented on the screen by the Visualization Manager. Queries formulated on the workbench are parsed and optimized for target architectures by the Query Manager, and then passed onto the execution engines. These can be either parallel or serial supercomputers, such as IBM SP1 and Intel Paragon supercomputers, single workstations, or workstation farms. The simplest queries consist of the extraction of well-defined features from "raw" data, without reference to any other information. These features are registered with the Information Depository to act as indices for further queries. Salient information extracted by queries can also be displayed via the Visualization Manager. The latter is implemented on top of IDL, and supports static plotting (2D and 3D graphs) of data, analysis of data (e.g., statistical, contours), and animation of datasets. Further details of the system architecture and an outline of typical querying sessions can be found in (Mesrobian et al. 1994).

Datasets

CONQUEST has been applied in the first instance to datasets obtained from two different sources. The first dataset is output from an Atmospheric Global Circulation Model developed at UCLA, chosen for two principal reasons: (1) it includes a challenging set of spatial-temporal patterns (e.g., cyclones, hurricanes, fronts, and blocking events); and (2) it is generally free of incomplete, noisy, or contradictary information. Hence it

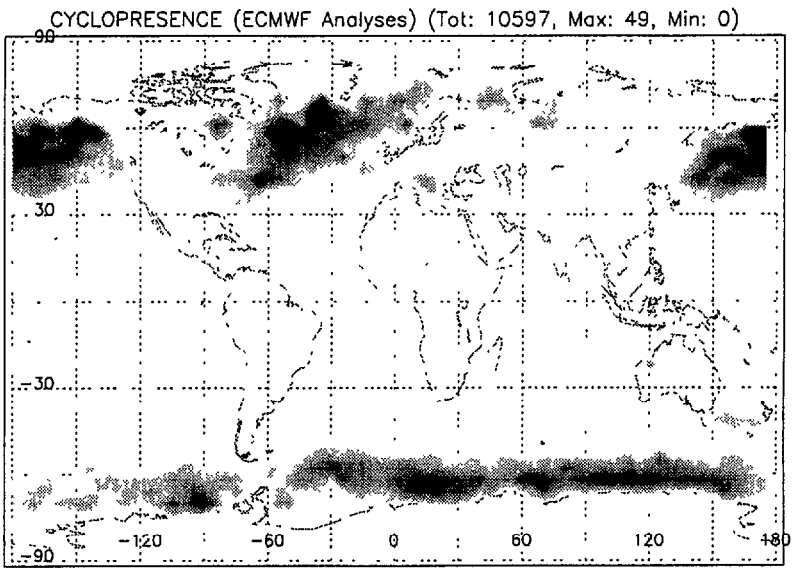

Figure 3: Cyclopresence density map of cyclones during the northern winter extracted from the ECMWF Analyses dataset (1985-1994).

mum is found by locating a grid location whose pressure value is lower than that at all the grid points in a neighborhood around the location by some (adjustable) prescribed threshold. This minimum is then refined by interpolation using low-order polynomials such as bi-cubic splines or quadratic bowls. Given a local minimum occurring in a certain GCM frame, the central idea is to locate a cyclone track by detecting in the subsequent GCM frame a new local minimum which is "sufficiently close" to the current one. Two minima are deemed "sufficiently close" to be part of the same cyclone track if they occur within 1/2 a grid spacing of each other. Failing this condition, they are also "sufficiently close" if their relative positions are consistent with the instantaneous wind velocity in the region. A trail of several such points computed from a series of successive frames constitutes a cyclone.

Figure 3 presents a cyclopresence density map of cyclones during the northern winter extracted from satellite ECMWF datasets. In the figure, white represents the lowest density value, while black indicate the largest density value. It can be seen that most extratropical cyclones are formed and migrate within a few zonally-elongated regions (i.e., "stormtracks") in the northern Atlantic and Pacific and off around the Antarctic.

Blocking Feature extraction

On time scales of one to two weeks the atmosphere occasionally manifests features which have well-defined structures and exist for an extended period of time essentially unchanged in form. Such structures are referred to, in general, as "persistent anomalies". One particular class of persistent anomalies, in which the basic westerly jet stream in mid-latitudes is split into two branches, has traditionally been referred to as "blocking" events. The typical anomalies in surface weather (i.e., temperature and precipitation) associated with blocking events and their observed frequency have made predicting their onset and decay a high priority for medium-range (5-15 day) weather forecasters.

While there is no general agreement on how to objectively define blocking events, most definitions require that the following conditions exist: 1) the basic westerly wind flow is split into two branches, 2) a large positive geopotential height anomaly is present downstream of the split, and 3) the pattern persists with recognizable continuity for at least 5 days. We modeled our blocking analysis operators after Nakamura and Wallace (Nakamura & Wallace 1990). Blocking features are determined by measuring the difference between the geopotential height at a given time of year and the climatological mean at that time of year averaged over the entire time range of the dataset. Before taking this difference, the geopotential height is first passed through a low-pass temporal filter (a 4th order Butterworth filter with a 6-day cut-off), to ensure that blocking signatures are not contaminated by the signals of migratory cyclones and anticyclones. The filtered field is averaged to obtain the mean year. A Fourier transform of the mean year is then taken, followed by an inverse Fourier transform on the first four Fourier components. This procedure yields smooth time series for seasonal cycles even if the dataset is small ($<\approx$ 100 years). The resulting filtered mean year is subsequently compared with the Butterworth-

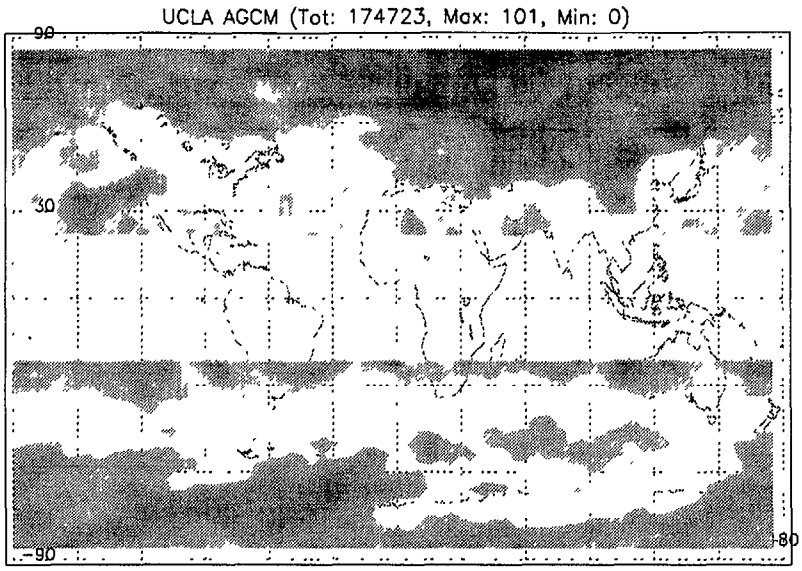

Figure 4: Density map of blocking events extracted from the UCLA AGCM model data-(1985-1989).

processed geopotential height fields to generate the fundamental anomaly fields. Blocking "events" can be detected as time periods δt during which filtered geopotential anomaly values are persistently higher than θ. Figure 4 presents a density plot indicating the global occurrances of blocking events for UCLA AGCM data (1985-1989), extracted using $\delta t = 5$ days and $\theta = 0.5\sigma$. In the figure, white represents the lowest density value, while black indicates the largest density value. Since blocking is by nature an extratropical phenomenon, we have eliminated values in the tropics from the plot.

Hierarchical Cluster Analysis

One method of reducing the dimensionality of the large image datasets, while retaining the structure of important regularities, is to perform some sort of cluster analysis which groups images together according to shared spatial features. We have adapted a hierarchical clustering procedure that has been used in the atmospheric science literature (Cheng & Wallace 1993), to a distributed farm of workstations. Our implementation uses publically available PVM message-passing software.

The procedure used by Wallace seeks to cluster together many frames of a geophysical field of interest generated over time. It builds clusters recursively by starting with each of N frames as an individual cluster. A Euclidean pointwise distance d_{pq} between every pair of images p amd q is first computed. The sum of all such distances is the defined as the error of a clustering. At each step two clusters are chosen to merged, namely that cluster pair which minimally increases the error. A simple mean image is updated at each step for each cluster for use in future computations of the total error. The resulting tree structure can be used to reduce the data dimensionality by identifying "cluster images" containing most of the important information. Many of the insights obtained correspond to similar observations that can be made on the basis of a somewhat more elaborate singular value decomposition analysis (Cheng & Wallace 1993).

Massively Parallel Feature Detection

The algorithms described above for extracting cyclone and blocking features on a 10-year dataset of atmospheric data require several hours to execute on a typical scientific workstation. Since one of our primary considerations is the need to supply an interactive facility, these processing times must obviously be drastically reduced if queries of this type are to supply initial indices to the datasets. Pre-processing and storage of indices by workstations is of course a feasible alternative for heavily used features, but will not suffice for a more general and wide-ranging querying capability. It is here that massively parallel processors (MPP's) enter the picture. The features described above can be computed quite efficiently on MPP's, bringing the turn-around time for a typical query down to the range of minutes on medium-scale parallel machines that have been used to date (a 24-node IBM SP1 and a 56-node Intel Paragon). It is expected that near real-time performance will be achieved when the system is ported to larger platforms comprising up to 512 nodes.

The parallel implementation of these queries requires an explicit decomposition of the problem across the various nodes of a parallel machine. This is by no means always a trivial task. In the case of cyclone detection, the optimal decomposition is based upon a

division of the problem into separate temporal slices, each of which is assigned to a separate node of the machine. A temporal decomposition such as this proves to be highly efficient on a coarse-grained architecture, provided that cyclone results obtained during a given time zone do not interfere too strongly with those at a later time.

Care must be exercised in such a decomposition, as the temporal dimension does not typically parallelize in a natural way, especially when state information plays an important role in the global result. State-information plays a fundamental role in the very definition of cyclones, so care must obviously be taken in the ensuing parallel decomposition. The problem proved tractable in the case of cyclone detection because of the observation that no cyclones last longer than 24 frames. This allows the use of a straightforward temporal shadowing procedure, in which each node is assigned a small number of extra temporal frames that overlap with the first few frames assigned to its successor node (Stolorz et al. 1995). In the case of blocking feature detection, a straightforward spatial decomposition which assigned different blocks of grid points to different machine nodes proves to be optimal. This type of decomposition has also be used for efficient parallel implementation of the hierarchical clustering algorithm.

Conclusions

We have outlined the development of an extensible query processing system in which scientists can easily construct content-based queries. It allows important features present in geophysical datasets to be extracted and catalogued efficiently. The utility of the system has been demonstrated by its application to the extraction of cyclone tracks and blocking events from both observational and simulated datasets on the order of gigabytes in size. The system has been implemented on medium-scale parallel platforms and on workstation farms.

The prototype system desribed here is being extended and generalized in several directions. One is the population of the query set with a wider range of phenomena including oceanographic as well as atmospheric queries. Another is the application of machine learning methods to extract previously unsuspected patterns of interest. A third issue is the scaling of system size onto massively parallel platforms, a necessary ingredient for the system to cope with the terabyte size datasets that are becoming available. In this regime, scalable I/O considerations are at least as important as those associated with computation *per se*, and are an active area of research. A final issue is the development of an appropriate field-model language capable of expressing queries based upon large imagery datasets rapidly and efficiently. Preliminary results (Skek and Muntz, private commmunication) suggest that the overhead introduced by the proposed language is relatively small (roughly 10%) compared to the execution of standalone code. These results are extremely encouraging, as the ability to formulate and add new queries to the system quickly and easily is of paramount importance to its usefulness as a scientific analysis tool.

Acknowledgements

This research was supported by the NASA HPCC program, under grants #NAG 5-2224 and #NAG 5-2225.

References

Knuth, E. and Wagner, L.M. 1992. *Visual Database Systems,* North Holland.

Chang, S-K. and Hsu, A. 1992. Image Information Systems: Where do we go from here?. *IEEE Transactions on Knowledge and Data Engineering* 4(5):431–442.

Niblack, W. et al. 1993. The QUBIC Project: Querying Images by Content Using Color, Texture, and Shape, IBM Research Division, Research Report #RJ 9203.

Fayyad, U. M.; Smyth, P.; Weir, N.; and Djorgovski, S. 1994. Automated analysis and exploration of large image databases: results, progress, and challenges. *Journal of Intelligent Information Systems* 4:1–19.

Gupta, A.; Weymouth, T.; and Jain, R. 1991. Semantic Queries with Pictures: The VYMSYS Model, in Proceedings of VLDB, 69–79. Barcelona, Spain.

Guptill, A. and Stonebraker, M. 1992. The Sequoia 2000 Approach to Managing Large Spatial Object Databases, in Proc. 5th Int'l. Symposium on Spatial Data Handling, 642–651, Charleston, S.C.

Mesrobian, E.; Muntz, R. R.; Santos, J. R.; Shek, E. C.; Mechoso, C. R.; Farrara, J. D.; and Stolorz, P. 1994. Extracting Spatio-Temporal Patterns from Geoscience Datasets, in IEEE Workshop on Visualization and Machine Vision, Seattle, Washington: IEEE Computer Society.

Stolorz, P.; Mesrobian, E.; Muntz, R. R.; Santos, J. R.; Shek, E. C.; Mechoso, C. R.; and Farrara, J. D. Spatio-Temporal Data Mining on MPP's 1995. submitted to Science Information and Data Compression Workshop, Greenbelt, MD.

Murray, R.J. and Simmonds, I. 1991. A numerical scheme for tracking cyclone centres from digital data. Part I: development and operation of the scheme *Aust. Met. Mag* 39:155–166.

Nakamura, H. and Wallace, J.M. 1990. Observed Changes in Baroclinic Wave Activity during the Life Cycles of Low-frequency Circulation Anomalies *J. Atmos. Sci.* 47:1100–1116.

Cheng, X. and Wallace, J.M. 1993. Cluster Analysis of the Northern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns *J. Atmos. Sci.* 50:2674–2696.