type." In *IEEE Transaction on Knowledge and Data Engineering*, Vol. 2, No. 1, 1990.

[4] A. Gupta, T. Weymouth, and R. Jain, "Semantic Queries with Pictures: The VYMSYS Model", *Proc. VLDB*, Barcelona, Spain, pp. 69-79, Sept. 1991.

[5] A. Guptill and M. Stonebraker, "The Sequoia 2000 Approach to Managing Large Spatial Object Databases", *Proc. 5th Int'l. Symposium on Spatial Data Handling*, Charleston, S.C., pp. 642-651, Aug. 1992.

[6] *IDL User's Guide*, Publication of Research Systems, Inc., Oct. 1991.

[7] E. Knuth, L.M. Wagner, Eds., *Visual Database Systems*, North Holland, 1992.

[8] C. R. Mechoso, C. C. Ma, J. D. Farrara, and J. A. Spahr, "Simulations of Interannual Variability with a Coupled Atmosphere-Ocean General Circulation Model." In *the Fifth Conference on Climate Variations*, American Meteorology Society, Boston, MA, 1991.

[9] C. R. Mechoso, C. C. Ma, J. D. Farrara, J. A. Spahr, and R. W. Moore, "Parallelization and distribution of a coupled atmosphere-ocean general circulation model." *Monthly Weather Review*, Vol 121, pp. 2062-2076, 1993.

[10] R. Muntz, L. Alkalaj, D. McCleese, C. Mechoso, J. Skrzypek and C. Zaniolo. "Data Analysis and Knowledge Discovery in Geophysical Databases", NRA-92-OSSA-2.

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

[12] W. Niblack, R. Barber, W. Equitz, et. al., "The QUBIC Project: Querying Images by Content Using Color, Texture, and Shape", IBM Research Division, Research Report #RJ 9203, February 1993.

[13] H. Samet, et. al., *The QUILT System (Version 3.0)*, Publication of the Department of Computer Science, the Center for Automation Research and the Institute for Advances Computer Studies, University of Maryland, 1990.

[14] M. Stonebraker, and G. Kemmitz, "The Postgres Next-Generation Database System", *CACM*, pp. 78-92, October 1991.

[15] M. Stonebraker, "An Overview of the Sequoia 2000 Project," *Proceedings 1992 COMPCON Conference*, San Francisco, CA, February, 1992.

[16] M. Darnovsky and J. Bowman, Transact-SQL User's Guide for SYBASE SQL Server, (Release 4.8), Publication of Sybase Inc., 1991.

[17] H. Le Treut, and E. Kalnay, "Comparison of observed and simulated cyclone frequency distribution as determined by an objective method", *Atmosfera*, Vol. 3, pp. 57-71, 1990.
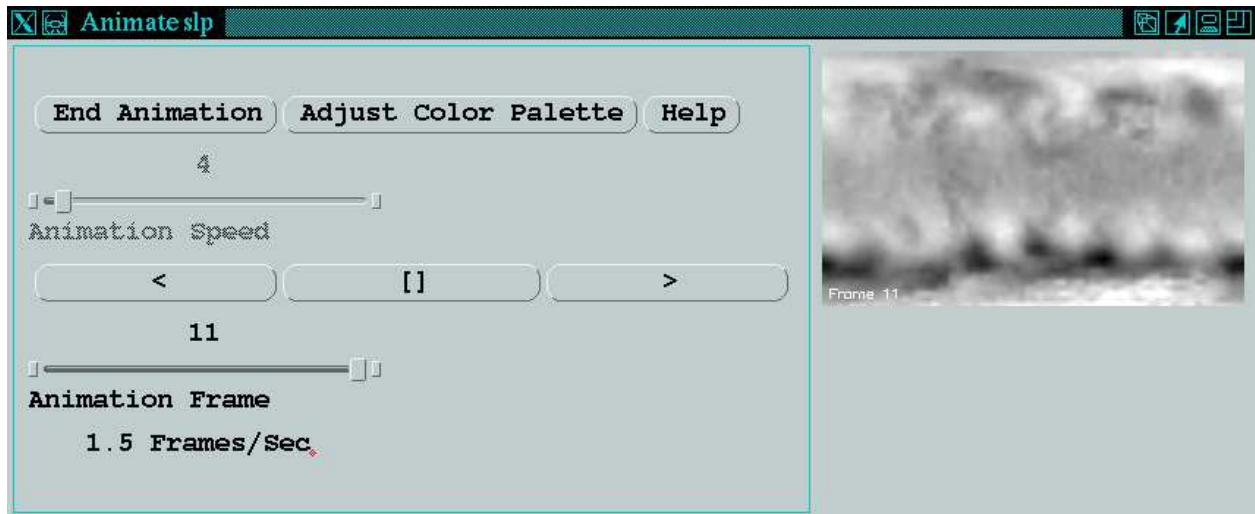
Figure 10: QUEST visualization manager animation of SLP fields during the life of the cyclone track shown in Figure 9.

6 Conclusion

Currently, scientists have access to large repositories of observational and model data, but lack high level analysis tools with which to investigate the data. Rather than being able to issue queries such as "Show me motion sequences that include hurricane (cyclone formation or temperature gradient patterns) that developed in the Southern Hemisphere", scientists are forced to issue queries like "Retrieve all temperature records produced between January 1, 1980 and December 25, 1991".

We have developed a prototype system called QUEST to provide content-based query access to massive datasets. To demonstrate the utility of spatial-temporal features as high-level indexes into terabyte datasets, we presented an algorithm for extracting cyclone tracks from sea level pressure data generated by a AGCM. A user scenario was presented to illustrate the interaction between a scientist and QUEST beginning with a query to select a subset of cyclones, and ending with the visualization of sea level pressure fields associated with the selected cyclone.

Much work remains to be performed to extend the functionality of QUEST. Work is underway to employ massively parallel processors such as the Intel Paragon, to extract spatio-temporal patterns from existing geoscience datasets, and to extract these features on-line as the model is executing. On-line extraction of features also provides the ability to computationally steer the model's execution in the event that desired features are not being produced by the model. A final area of current research is the use of Distributed Object Management Systems (DOMS) to provide support for integrating data analysis, visualization and data management in a heterogeneous distributed environment.

Acknowledgements

References

[1] A. Arakawa, and V. R. Lamb, "Computational Design of the Basic Dynamic Processes of the UCLA General Circulation Model." *Methods in Computational Physics*, Vol. 17, Academic Press, 1977.

[2] S-K. Chang, and A. Hsu, "Image Information Systems: Where do we go from here?", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 4, No. 5, pp. 431-442, Oct. 1992.

[3] D. Chimenti, R. Gamboaa, R. Krishnamurthy, S. Naqvi, and C. Zaniolo, "The $\mathcal{LDL}$ System Proto-
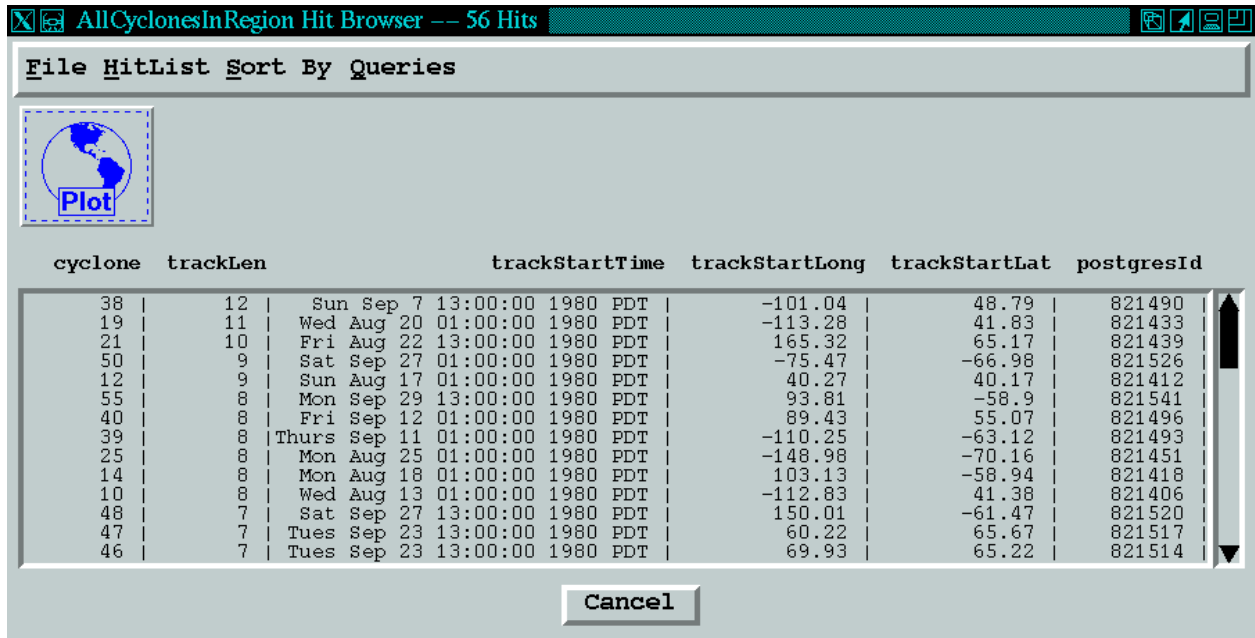
AllCyclonesInRegion Hit Browser -- 56 Hits

File HitList Sort By Queries

Plot

cyclone	trackLen	trackStartTime	trackStartLong	trackStartLat	postgresId
38	12	Sun Sep 7 13:00:00 1980 PDT	-101.04	48.79	821490
19	11	Wed Aug 20 01:00:00 1980 PDT	-113.28	41.83	821433
21	10	Fri Aug 22 13:00:00 1980 PDT	165.32	65.17	821439
50	9	Sat Sep 27 01:00:00 1980 PDT	-75.47	-66.98	821526
12	9	Sun Aug 17 01:00:00 1980 PDT	40.27	40.17	821412
55	8	Mon Sep 29 13:00:00 1980 PDT	93.81	-58.9	821541
40	8	Fri Sep 12 01:00:00 1980 PDT	89.43	55.07	821496
39	8	Thurs Sep 11 01:00:00 1980 PDT	-110.25	-63.12	821493
25	8	Mon Aug 25 01:00:00 1980 PDT	-148.98	-70.16	821451
14	8	Mon Aug 18 01:00:00 1980 PDT	103.13	-58.94	821418
10	8	Wed Aug 13 01:00:00 1980 PDT	-112.83	41.38	821406
48	7	Sat Sep 27 13:00:00 1980 PDT	150.01	-61.47	821520
47	7	Tues Sep 23 13:00:00 1980 PDT	60.22	65.67	821517
46	7	Tues Sep 23 13:00:00 1980 PDT	69.93	65.22	821514

Cancel

Figure 8: QUEST query hit browser.

Quest IDL Tracks

Close
Restore
ZoomIn
Animate2
Animate

Bounding Box
LON1 = -116
LAT1 = 64
LON2 = -45
LAT2 = 6
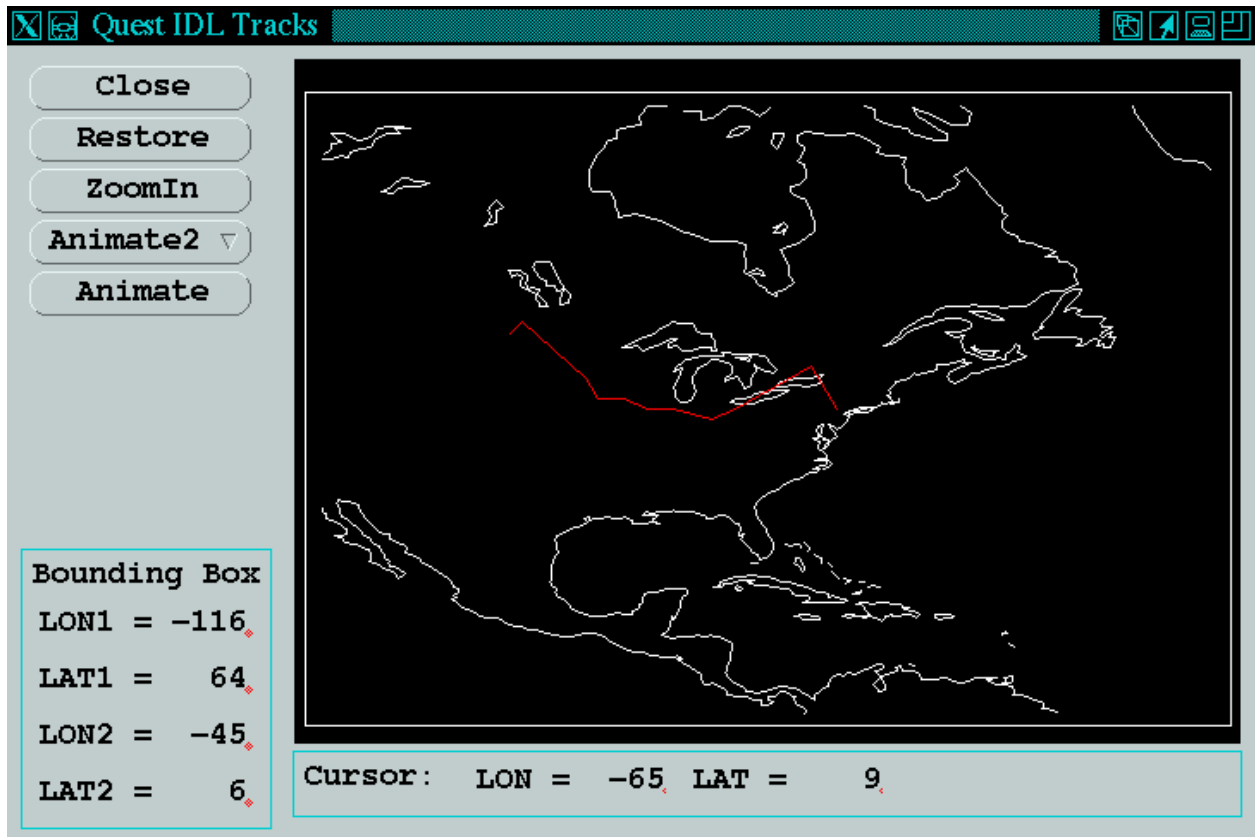
Cursor: LON = -65 LAT = 9

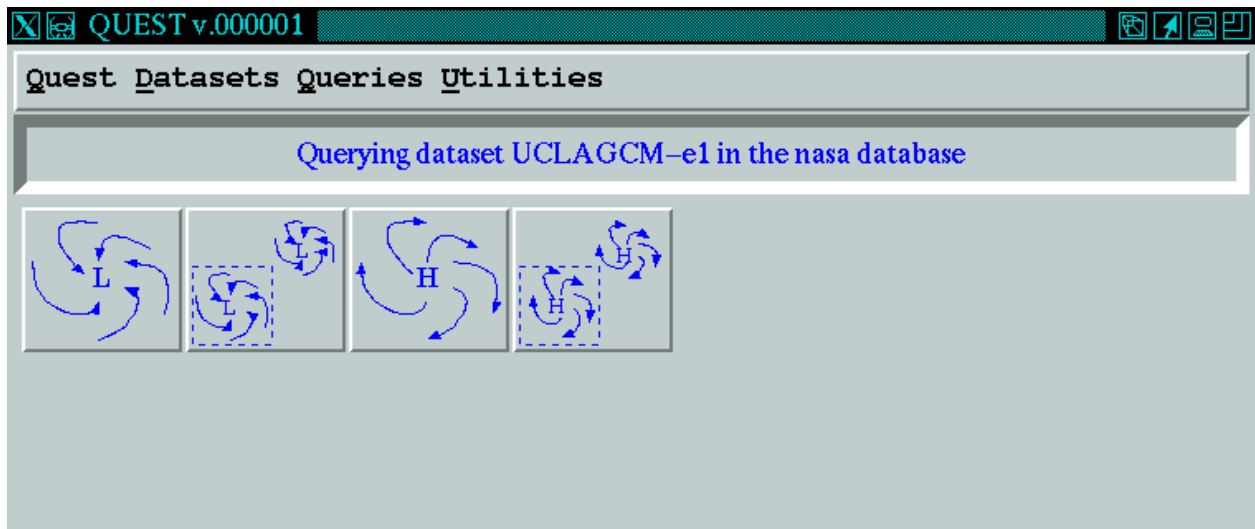Figure 9: QUEST visualization manager plot of a selected cyclone track.

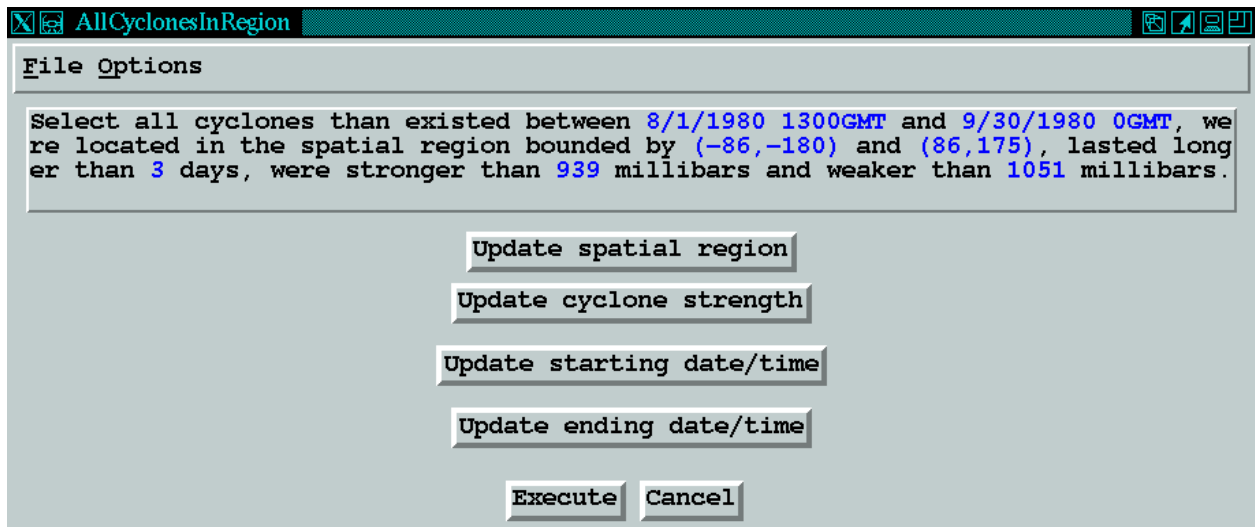Figure 6: QUEST GUI for selecting a feature query icon.

Figure 7: QUEST panel for specifying the spatial, temporal, and magnitude extents for desired cyclones.
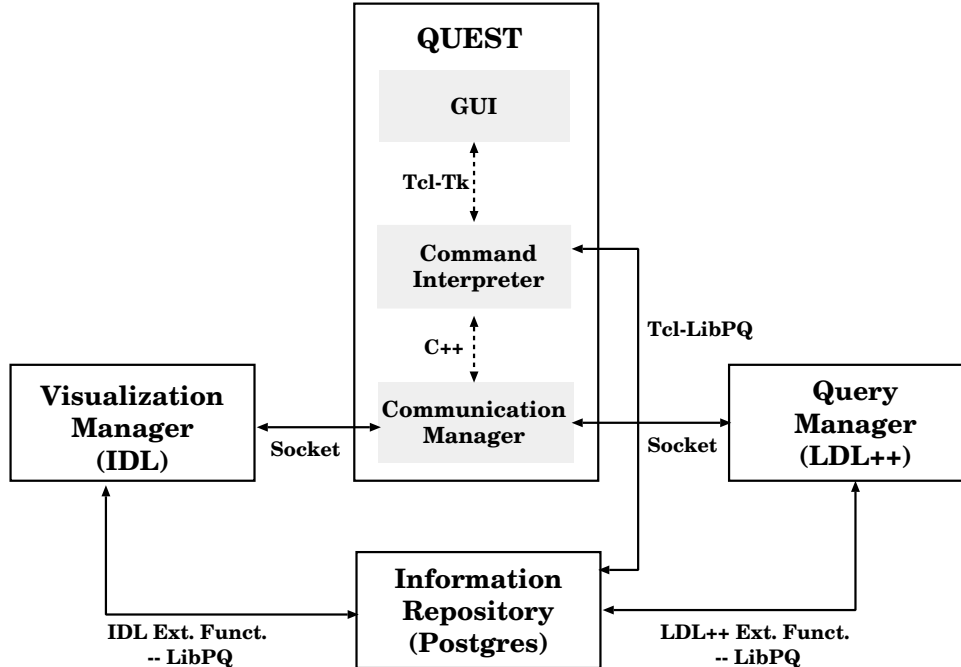
Figure 5: QUEST system architecture.

to request additional information (via QUEST or the information repository).

The graphical user interface (GUI) provides the ability to query scientific data and extracted features in the information repository. The GUI was built using the Tool Command Language (Tcl) and X11 Toolkit (Tk) developed at UC Berkeley. The GUI construction is highly interactive since GUI's schema is stored in the information repository.

## 5 Content-based access to datasets using extracted features

Once features have been extracted from a large dataset, they can used as high-level indexes into the dataset. This ability is very appealing when datasets sizes are in the gigabyte or terabyte range, and the only access method is simply based on spatial and temporal coordinates.

To illustrate the use of the extracted features, consider the QUEST GUI panel shown in Figure 6. The icons represent the various high-level features that are available as indexes into the dataset being investigated. The first two icons are for querying the system about all cyclones or cyclones that existed over a given spatio-temporal interval, while the latter two provide similar support for anticyclones. Suppose that

the user wants to query the system to locate cyclones of interest (this is the first step to accessing the data subset to be analyzed and/or visualized). The user query form shown in Figure 7 is automatically generated by QUEST when the second icon is clicked.

The scientist can specify the desired spatial coordinates (extent), the temporal interval, and the minimum and maximum SLP values for cyclone minima, in an attempt to limit the database search. QUEST transforms this graphical query into its internal query language and communicates with the $\mathcal{LDL}++$ query manager to process the query. The query manager then interfaces with the repositories (e.g., Postgres, Sybase) to retrieve the requested information. Figure 8 presents the "hits" (i.e., the cyclone tracks that match the user-specified criteria). The scientist can sort the hits, and select the subset of hits to be further investigated (e.g., plotted, etc.).

To plot a set of cyclone tracks, a scientist simply clicks on the plot icon. QUEST communicates with the repositories to retrieve the tracks, and passes the information along to the visualization manager, which in turn produces the plot (see Figure 9). A scientist can then interact with the plot window to further select a subset of tracks, and can visualize animations of model variables during the life of the selected cyclone tracks (see Figure 10).
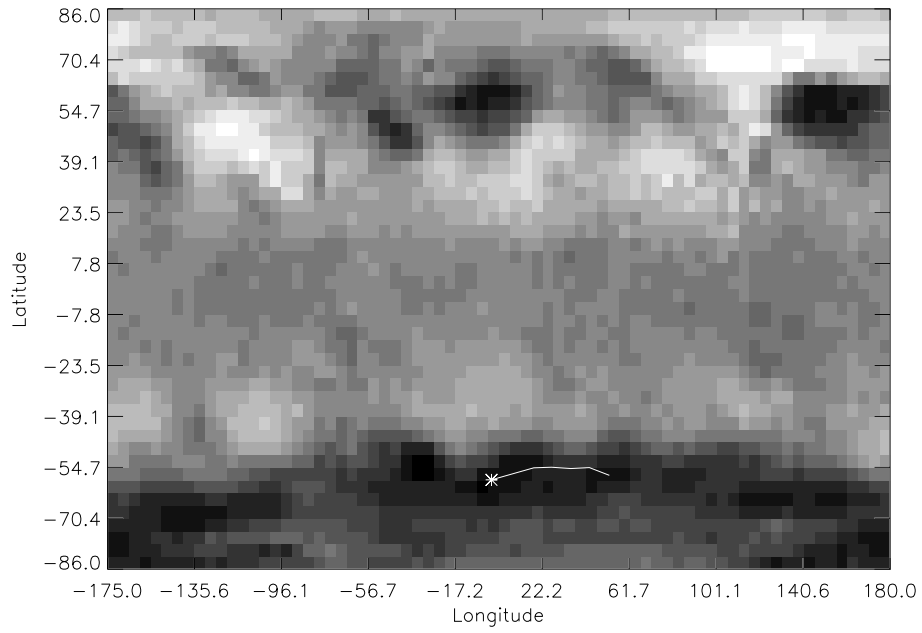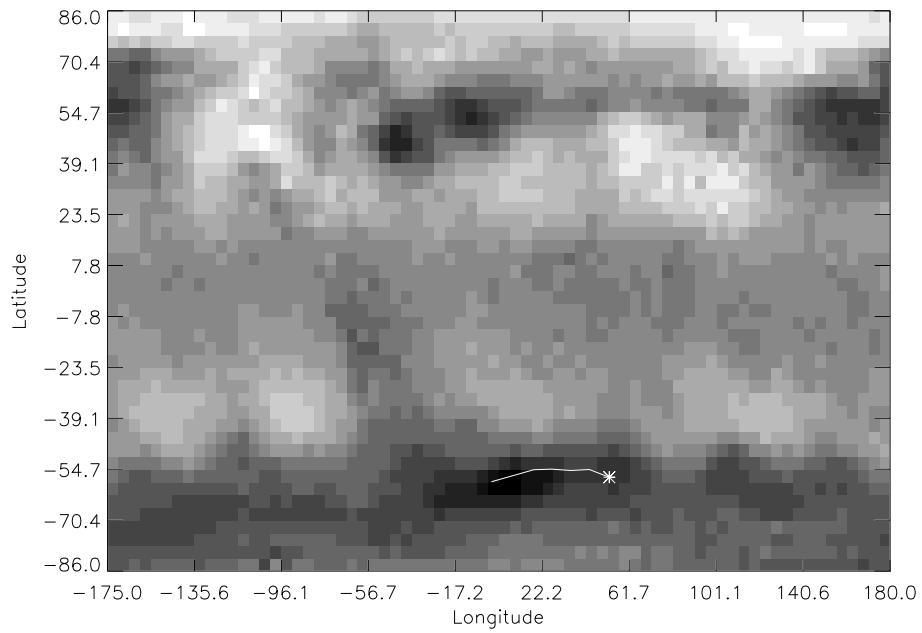
Figure 4: Plots of sea level pressure fields overlaid with cyclone track (extracted minimas are represented by asterisks): (a) Start of a particular cyclone track on January 3, 1981, 1200GMT; (b) End of a the same cyclone track on January 6, 1981, 0GMT.

considered to be a part of the same cyclone track if they satisfy one of the following two criteria:

- The difference between the spatial coordinates of min^t and min^{t+1} is less than within a 1/2 grid cell (in our study, 2.5° in East-West direction and less than 2.0° in North-South direction). Formally, the criterion can be expressed as $dist_x(min^t, min^{t+1}) < 2.5°$ *and* $dist_y(min^t, min^{t+1}) < 2.0°$.

- The location of the two minima is consistent with the direction and velocity of the wind at the 700 millibar level (based on the assumption that wind and cyclone tend to move in the same direction). In other words, the min^{t+1} is located in the direction indicated by the 700mb wind variable sampled at the location of min^t. Given that the direction is correct, the distance between the minima is less than the distance that could be traveled given the wind velocity. Typically, the maximum distance that could be traveled by a cyclone between t and $t + 1$ was 1 grid cell.

A final criterion on a cyclone track is that it must last for at least 3 simulated days. This criterion greatly reduces the number of cyclone tracks detected and stored in the database. To demonstrate the performance of the tracking algorithm, consider the SLP fields shown in Figure 2. The two SLP fields correspond to an elapsed time of 3 simulated days. Figure 4 presents an example cyclone track that was extracted during that period. The cyclone began on Jan 3, 1981 at 1200GMT at the minima marked by the asterisk (see Figure 4(a)). The entire track has been overlaid to indicate the course taken during the cyclones life. Figure 4(b) presents the SLP field and the location where the cyclone track ended.

To test our prototype system, we have ten years worth of selected fields from GCM model data (approximately 2 Gbytes). The extraction of 3686 cyclone tracks required the analysis of 264 Mbytes (SLP fields plus wind velocities at 700 millibars), and took about an hour per model year on a Sun Sparc 10/30. We are currently implementing a parallel version of the the algorithm on an 64 node Intel Paragon. Preliminary results suggest that we can anticipate to process a model year on the order of minutes.

The algorithm described above is similar to others proposed in the literature. Le Treut and Kanaly [17] extracted minima by computing a mean over a cell's closet 20 grid points and accepting those points where the difference between the SLP value at the point was 4 millibars lower than the mean. A cyclone track is formed by comparing minima in a given test area (20° × 25°) with minima that existed during the previous time step. If more than one minima exists in the test area, wind information at 500 millibars is used to select the cyclone that is heading in the direction indicated by the wind information. A cyclone track is kept if it lasts for at least 3 time steps. Murray and Simmonds' [11] tracking algorithm computes estimates of the new location of a min^t, by calculating the probability of associations between the predicted and realized positions for each minima, and selects the matching of these associations with the highest overall probability (i.e., credit assignment).

4 System architecture

Figure 5 presents our environment for scientific data analysis, knowledge discovery, and visualization. The QUEST information system comprises a graphical user interface, a query manager, visualization manager, and an information repository. It should be noted that these components run as separate processes and communicate with each other using the protocols appearing on the communication links.

The query manager maintains a unified schema of the data stored in distributed, heterogeneous repositories (Postgres [14], Sybase [16], QUILT [13]), and provides QUEST with a common query language (application interface) to the underlying repositories. We implemented our query manager using $\mathcal{LDL}$++, a deductive DBMS developed at MCC [3]. $\mathcal{LDL}$++ provides a logic-based query language, complex data types (e.g., sets, lists, composite types), and an interpretive environment for rapid prototyping. We extended $\mathcal{LDL}$++ to support spatial and temporal data types (along with operators).

Our initial information repository was built using Postgres, an extensible relational DBMS. Postgres provides inheritance, abstract datatypes, user-defined functions, and large object (blob) support. We defined classes for our spatial-temporal objects (such as cyclone tracks) as well as objects such as model data datasets, observation datasets and experiments.

The visualization manager supports static plotting (2D and 3D graphs) of data, analysis of data (e.g., statistical, contours), and animation of datasets. We chose to implement our visualization manager on top of IDL. We augmented IDL with a communication module and support for maintaining object ids of data being manipulated. These object ids are used by IDL
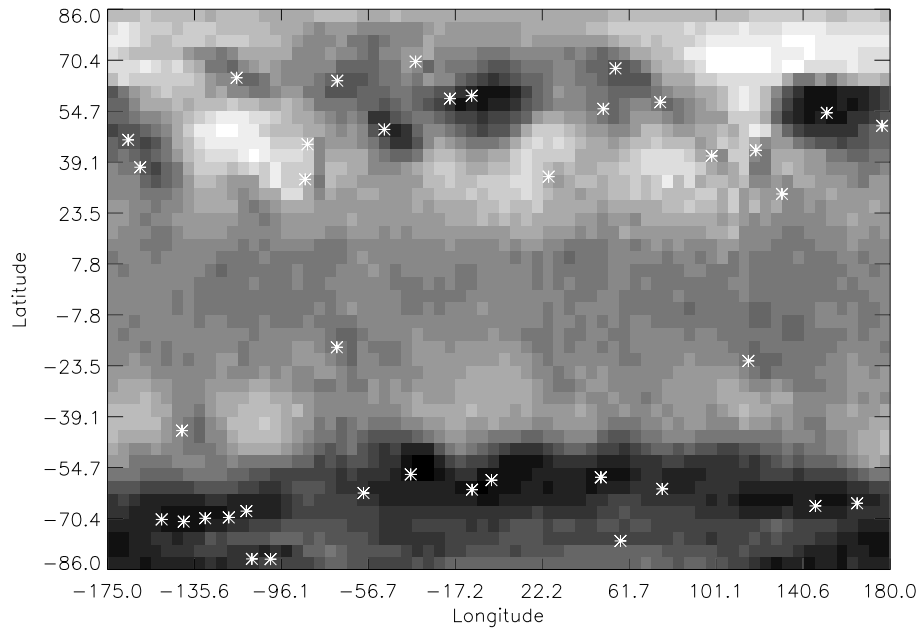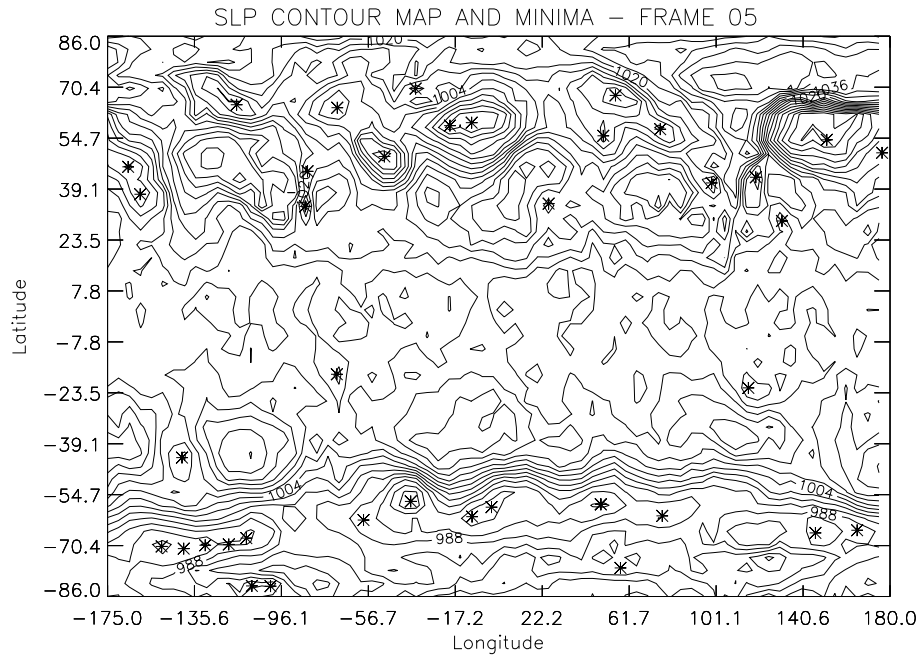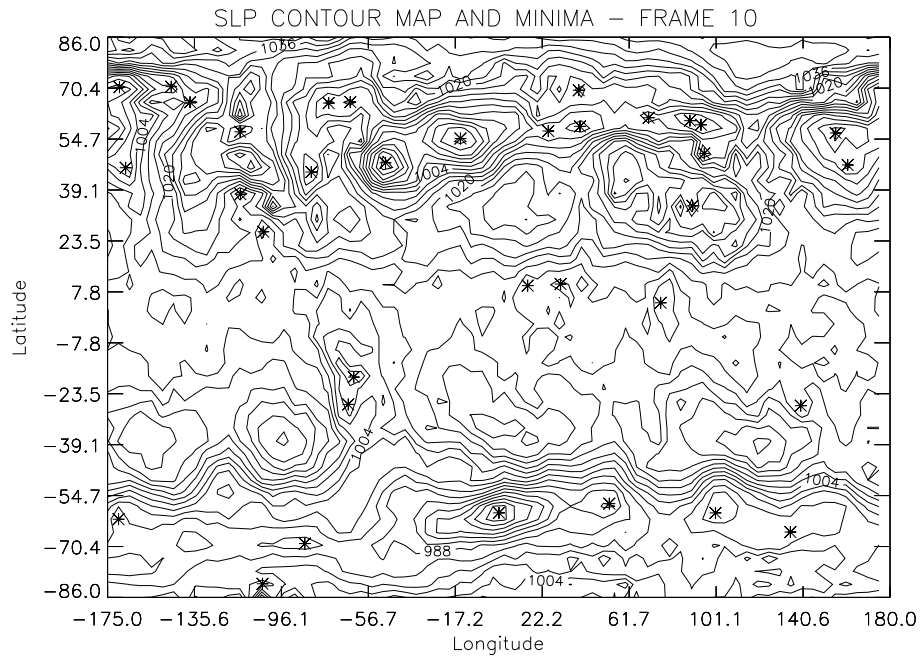
(a)
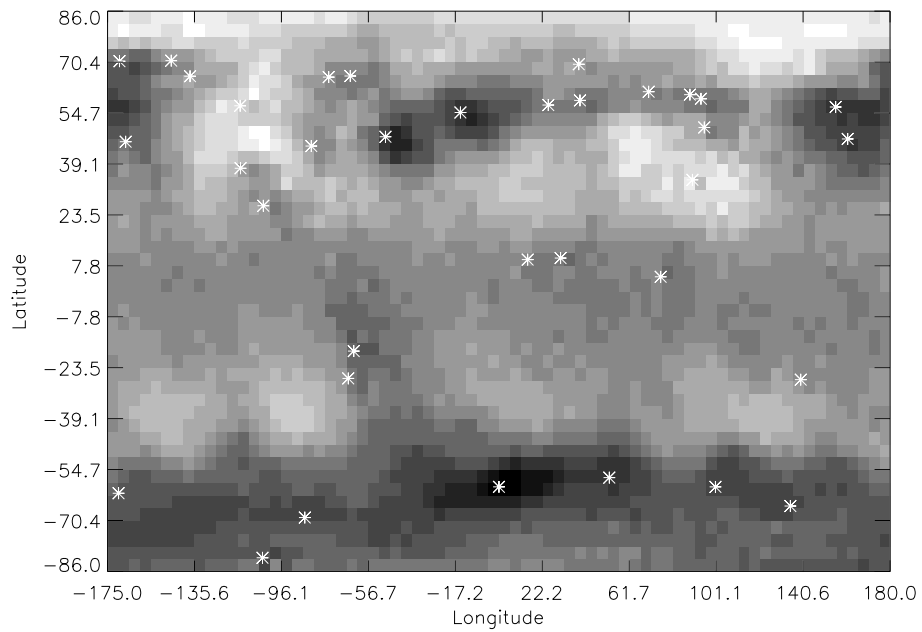
(b)

Figure 2: Plots of sea level pressure fields (SLP) overlaid with extracted minima (asterisks): (a) SLP field for January 3, 1981, 1200GMT; (b) SLP field for January 6, 1981, 0GMT.

Figure 1: Contour plots of sea level pressure (SLP) fields overlaid with extracted minima (asterisks): (a) SLP field for January 3, 1981, 1200GMT; (b) SLP field for January 6, 1981, 0GMT.

The prognostic variables of the AGCM are horizontal velocities, potential temperature, water vapor and ozone mixing ratio, surface pressure, ground temperature, and depth of the planetary boundary layer. There are also diagnostic variables such as vertical velocities, precipitation, cloudiness, surface fluxes of sensible and latent heat, surface wind stress and radiative heating. Typically, the model's output is written out to the database at 12-hour (simulation time) intervals; however, this frequency can be modified depending on storage capacity of the database. the model can be run with different spatial resolutions (grid sizes) and temporal resolution (output frequency). At the lowest spatial resolution ($4° \times 5°$, 9 levels) with 12 hour output interval, the AGCM produces approximately 5 Gbytes of data per simulated year, while a 100-year simulation of a AGCM with a $1° \times 1.25°$, 57 levels) generates approximately 30 terabytes of output.

3   Extracting spatio-temporal features

We are interested in capturing features (usually spatial) and tracking them over time. Phenomena simulated by the AGCM include extratropical cyclones, blocking events, hurricanes and fronts. In this paper we focus on the design of an algorithm for the detection and monitoring of cyclone tracks. In a weather map, the signature of an extratropical cyclone is a set of closed contours surrounding a minimum in sea level pressure (SLP). The global distribution of sea level pressure is provided by the AGCM at regular time intervals. As time advances, a cyclone may translate in space along a trajectory also called a "cyclone track".

To help visualize the location of cyclone centers, Figure 1 presents contour plots overlaid with extracted local minima (asterisks) for two SLP fields. It should be noted that not all of the extracted minima shown in Figures 1(a-b) are true cyclone centers and consequently a part of a cyclone track. Figure 2 presents plots of the SLP fields overlaid with their respective minima.

Given a time sequence of SLP fields (with a particular temporal separation), we developed an algorithm to extract cyclone tracks (see Figure 3). This algorithm consists of three steps: 1) minima extraction, 2) minima location refinement, and 3) assignment of minima to tracks.

In the first step, local minima in the SLP data field are located. A local minima exists at (x, y) if the corresponding SLP value satisfies the following criteria:

- SLP(x,y) is less than each one of the values for

```
for each frame
   compute all minima;
   for each minimum
      for all active tracks
         if the minimum satisfies closeness criteria
            include it on that track;
         endif
      endfor
      if the minimum was not added to any track
         create a new track and add
         the minimum to it
      endif
   endfor
   for all tracks that didn't get a new minimum
      if track is greater than $length_{min}$ (3 days)
         save it as valid track
      else
         discard track
      endif
   endfor
endfor
```

Figure 3: Algorithm for extracting cyclone tracks.

grid points in the immediate neighborhood of the point in question (i.e., 8-neighbor).

- SLP(x,y) is less than the average value computed using a 5x5 neighborhood centered at (x, y) (but not including the center point). The difference threshold was set to 5.5 millibars. This criteria permits the detection of large, shallow low pressure areas.

Since the spatial resolution of the AGCM output that we are using in this study is rather coarse ($4° \times 5°$), the location of extracted minima may not vary smoothly with time, further complicating cyclone tracking. To obtain a more continuous variation in the locations of pressure minima, we used a method developed by Murray and Simmonds [11]. In this method, the SLP field is fitted by a bicubic spline function. Gradients in the interpolated surface are then used to better locate the centers of the minima extracted in the first step.

Given an SLP field along with its interpolated minima, the final step in the algorithm is to try to associate each minima in the frame at time $t + 1$ with cyclone tracks that have been monitored during the previous time steps ($\ldots, t - 2, t - 1, t$). Two minima computed at successive frames (min^t and min^{t+1}) are

repository of contextual and semantic information.

c) The nature of exploratory data analysis for scientific hypothesis testing or phenomenon detection is basically an iterative, successive-refinement process. The scientist initially applies a coarse model on the data, and then uses the outcome of this first experiment to refine his/her model and methods; then the process is repeated until the hypothesis is dropped or it is refined into one that is fully corroborated by the collected data. For such investigations to be practical, the scientist must have at hand a powerful system that supports: (1) the easy formulation of powerful queries and discriminant decision rules against the database; (2) a natural representation of the relationships of the scientific domain of interest (e.g, in natural domains of the space and time, but, possibly, in the frequency domain as well); and (3) efficient execution of these queries without requiring the scientist to become cognizant of the storage structures and processing strategies involved.

d) Once methods for detecting patterns of interest have been established, the system can search for these patterns as new data is added from sensors and satellites, and through a trigger-based activation mechanism alerts interested scientists. Furthermore, since the system automatically records as metadata which datasets, algorithms and parameters were used in the experiments, the database becomes the companion logbook of each scientist. As scientific theories are revised and improved, the system will help scientists to revise results obtained under old assumptions (a giant make file).

Spatio-temporal analysis of dynamic events such as the motion of rigid or nonrigid bodies contributes to the image (data) understanding tasks by disambiguating scene information, whenever the observer and/or objects in the scene are moving. Information that can be extracted from motion cannot be obtained from any other attributes of the image. Image databases must solve the problem of querying dynamic processes if they are to be useful in retrieving information encoded in, for example, all sequences of "data" frames with circular cloud pattern that resembles a hurricane. At issue is how users can utilize distributed image motion databases, both in storing new sequences, retrieving existing ones, and comparing sequences for situation assessment purposes. We are interested in systems that can respond to queries of the type, "Show me motion sequences that include a hurricane (cloud formation or temperature gradient patterns) like the one in this sensory data".

Content based access to image databases has become an active area of research in recent years, but is still in its infancy. A sampling of recent work can be found in [7, 2]. The QUBIC project at IBM [12] is an example system that illustrates the state-of-the-art in image retrieval by content, while examples of work in the area of geoscience databases include VIMSYS [4] and Sequoia 2000 [5].

As part of a NASA HPCC Grand Challenge effort [Mun92], we have developed a prototype system called QUEST to provide content-based query access to massive datasets used in geophysical applications. QUEST employs workstations as well as teraFLOP computers to produce spatio-temporal features that are used as high-level indexes into terabyte datasets. Our first application area is the output of global change climate models and in the initial prototype, the first features extracted for content-based access are model-simulated trajectories of cyclones and anticyclones. This paper presents an algorithm for extracting cyclone trajectories from simulations performed with a General Circulation Model of the atmosphere (the UCLA AGCM), and illustrates the use of cyclone indexes to access (via QUEST) subsets of GCM data for further analysis and visualization.

2   Geophysical datasets

Geophysical datasets are generally produced by either observational systems (e.g., satellites) or models. Earth science phenomena, modeled or observed, typically contain features which can be extracted from their datasets. These spatio-temporal phenomena and their derived features can be managed, manipulated, and indexed by applications according to their spatial and temporal properties.

We chose the output of AGCMs as our application domain 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 contradicting information. Hence it serves as an ideal testbed for validating our prototype environment.

The specific data set used in this study was generated by the UCLA AGCM [1, 8, 9]. The horizontal structure of the model is based on grid cells of various resolutions; we are using a grid size of 5° longitude and 4° of latitude. The vertical component of the model is represented by a series of layers between the Earth's surface and a prescribed pressure level in the upper troposphere or stratosphere.

Extracting Spatio-Temporal Patterns from Geoscience Datasets

E. Mesrobian, R. R. Muntz
J. R. Santos, and E. C. Shek
UCLA Computer Science Dept.

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

P. Stolorz
Jet Propulsion Lab
CalTech

Abstract

A major challenge facing geophysical science today is the unavailability of high-level analysis tools with which to study the massive amount of data produced by sensors or long simulations of climate models. We have developed a prototype system called QUEST to provide content-based access to massive datasets. QUEST employs workstations as well as teraFLOP computers to analyze geoscience data to produce spatial-temporal features that can be used as high-level indexes. Our first application area is global change climate modeling. In the initial prototype, the first features extracted are cyclones trajectories from the output of multi-year climate simulations produced by a General Circulation Model. We present an algorithm for cyclone extraction and illustrate the use of cyclone indexes to access subsets of GCM data for further analysis and visualization.

1 Introduction

A critical challenge facing geophysical science today is the unavailability of high-level analysis tools with which to study the massive amount of information captured by sensors onboard orbiting satellites or produced by climate models. To address this challenge, we must develop a new generation of systems for scientific data management capable of coping with

- the highly complex multimodal queries and intelligent retrievals required for scientific investigations and knowledge discovery, and

- the staggering computational demands posed by the extraordinary size of the data set and the complexity of the tasks involved.

Both facets of this challenge are present in queries involving detection of changes and flow structures that are common in atmospheric and geophysical science.

Thus, the recognition of phenomena such as monsoons, extratropical cyclones, cold and warm fronts, ocean currents and ocean eddies is rich both in data complexity (large multidimensional data sets) and logical complexity (difficulty of detecting and tracking patterns evolving in a multidimensional space). These applications must be supported by (1) large parallel systems capable of providing fast access to very large data sets on mass storage and (2) scientific information management systems capable of taming the complexities of the logical task involved. In fact, scientific data management systems capable of supporting successful geophysical investigations on data products generated by models or by Earth-Observation Systems (EOS) must address several major issues:

a) There is a very wide gap between the high-level conceptual abstractions with which scientists operate (involving, e.g., trends, evolution and correlations) and the very low level at which data is collected (i.e., vectors of noisy data samples). Therefore, intermediate levels of derived data products are used to produce a dataset more conducive to scientific investigations (e.g., through enhancements and cleaning). Thus, there is a need for a vertically integrated architecture which manages the mapping between different product levels, to ensure complete relatability between levels (e.g., via metalevel data) and optimization of the high-level queries and goals of interest.

b) Previous experience with recognition of geophysical phenomena from observational data indicates that this cannot be performed by a mere (bottom-up) enhancement of the observational data followed by the recognition of the syntactic or geometric patterns of interest: successful recognition also requires extensive use of contextual data whereby a great deal of semantic information (e.g. about regions, time, and plausibility) is passed down to the lower layer of processing along with semantic expectations about the goals being sought. This suggests that successful exploration should be performed in a vertically integrated system, where data-driven and goal-directed processing are combined, and full advantage is taken of a rich