Visualization Benchmark Specification
for
Sequoia 2000 Project

PRELIMINARY DRAFT

DO NOT DISTRIBUTE

Mike Olson
Jim Frew
Peter Kochevar
Duane Waliser
Allison Woodruff

1. Introduction

This document presents a preliminary set of benchmarks for the visualization of scientific data by researchers participating in the Sequoia 2000 Project. This document is preliminary for the following reasons:

- The Sequoia 2000 Swat Team still plans a visit to the California Department of Water Resources, and will not issue a final report until after that.
- This draft is to be reviewed by Project members before adoption.
- This draft provides a schedule only for the implementation of preliminary features of the benchmark. The implementation deadline for the preliminary features is January 10, 1993. A more extensive benchmark is planned, and will be specified in more detail later.

The rest of this document is organized as follows. Section two presents a summary of three Sequoia 2000 pilot projects. Section three describes the data management requirements that these pilot projects impose on the underlying computer system. Section four lists some of the data manipulation tools that will be required to analyze and process the data generated by the projects. Section five specifies the services that must be provided by the visualization software to support the pilot projects. Section six lists other areas in which computer systems support is required, but for which none is planned before the January deadline. Section seven gives a detailed breakdown on the services that will be provided by January. Section eight summarizes additional commitments that have been made by Sequoia group members, and which will be met before January. Finally, the original Sequoia Swat Team trip reports are included as appendices to this document.

2. Summary of Pilot Projects

The following pilot projects were selected at the last Sequoia retreat to foster cooperation among researchers and the development of computer support for global change research.

2.1. Air/Sea Interaction

The goal of this project is to understand climate processes over the open ocean. This research includes

- finding an algorithm that separates snow and ice from cloud cover in Advanced Very High Resolution Radiometer (AVHRR) satellite images;
- energy absorption modelling in atmospheric and oceanic systems; and
- an analysis of heat transfer by the evaporation or condensation of water at the ocean's surface.

This project requires tools to map heterogeneous data to a single coordinate system, statistical analysis tools, and tools that allow users to explore relationships among data visually.

2.2. Tropical Seasonal Cycle

This project compares the output of climate models with observational data, including satellite images and ship traces. First, the data must be analyzed to identify periodic functions that drive variation. Second, the correlations among the component functions must be identified. Finally, experiments must be designed to test hypotheses that explain why these functions drive the climate system.

The problems confronting researchers in this area are the heterogeneity of their data, complex and non-standard analysis tools, and a lack of good visualization software for examining raw data and the results derived from it.

2.3. Detection of California Wildfires

AVHRR satellite images can be used to compute the "greenness" of pixels on land, giving a good indication of the amount of healthy vegetation they contain. Natural processes such as wildfires affect vegetation cover. This project attempts to detect wildfires from satellite imagery.

The project requires software for processing satellite images over time, and for displaying and animating these images and derived images.

3. Data Management Requirements

This section lists the data management requirements that are common to the pilot projects. These services are also required by Sequoia researchers working on other projects.

3.1. Storage

Sequoia researchers must be able to store large volumes of data conveniently. The Project owns roughly ten terabytes of disk and tape storage, and must provide access to it with software.

3.2. Schema

To foster cooperation among researchers, Sequoia must specify common access formats for scientific data. This specification must include both tabular data, stored in conventional database relations, and large object data, stored either in the database or in a file system.

3.3. Programmatic Interface to the Data Store

Programmers must be able to fetch tabular and large object data from the database system easily. Similarly, tools for manipulating data stored in a conventional file system are important. For example, users should be able to have a set of satellite images automatically staged onto their local disks for short-term processing, and have the space reclaimed automatically when the images are no longer needed.

4. Data Manipulation Requirements

Sequoia should provide tools that permit scientists to analyze data from many different sources. This section describes the most critical areas in which support is required.

4.1. Spatial Transformation

Scientific data acquired in different spatial coordinate systems must often be transformed to a common coordinate system prior to analysis. Sequoia Earth scientists would like Sequoia to develop a generic coordinate transformation tool, as they have found existing tools to be awkward, inconsistent, or erroneous.

Important features of the transformation tool include the ability to translate between common map projections and to interpolate irregularly-spaced data onto a regular grid.

4.2. Analysis

Sequoia must support sophisticated analysis of scientific data. The types of analyses common to Project researchers include

- Complex Empirical Orthogonal Functions (CEOF);
- analysis of variance (ANOVA);
- conventional statistical processing (computation of mean, variance, standard deviation, and correlation);
- field operations, such as adding or subtracting two images; and
- symbolic mathematics, including integration and differentiation of functions over some variable.

5. Visualization Requirements

Visualization is an important tool to permit scientists to explore relationships among data and patterns in data. The following features are important in a data visualization tool:

- It must be easy for scientists to use; assistance from a programmer must not be required.
- It must allow data from many different sources, including satellite images, observational data, and model output, to be viewed simultaneously.
- It must support visualization of three- and four-dimensional[1] data.
- It must support animation of sequences, zooming and panning over images, and interactive changes to the visualization to change its appearance.

[1] In the context of Sequoia, the dimensions of interest are nearly always three spatial and one temporal.

The visualization tools used most widely by Project researchers are IDL, IPW, and AVS. The Project is using AVS in particular as a prototype for the "Tioga" visualization interface to POSTGRES.

6. Additional Features

Data management, analysis, and visualization are required to support the pilot projects identified in Section Two. This section lists other features of interest to Project researchers. These features will be included in a future benchmark specification, but will not be supported by January 1993.

6.1. Data Lineage

Scientists must be able to annotate data as it is processed by Sequoia. The annotation must follow the data, so that annotations from raw input data are available to a scientist examining a product derived from it. In addition, the Sequoia database should automatically maintain the processing history of all the data it stores, allowing researchers to trace operations applied to the data in the past.

6.2. Computational Steering

Scientists must be able to intervene in the execution of a model or data analysis, changing variables or parameters to alter the outcome. This capability is called *computational steering*, and is not supported by the current set of models used by researchers. Sequoia should explore support for computational steering.

7. Benchmarks

This section lists services that will be provided by Sequoia by January 10, 1993. Completion dates and names of responsible Project members are included.

7.1. Data Management

By 31 December 1992:

- Support for access to 10 Terabytes of storage from the POSTGRES database system (Mike Olson, Randy Wang).
- Support for access to 10 Terabytes of storage via NFS (Mike Olson, Case Larsen, Randy Wang).

By 10 January 1993:

- Large-volume scientific data loaded onto the Bigfoot (10TB) data server (Duane Waliser, Norm Hall, Emelia Villaros-Bainto, Terry Figel, Zahid Ahmed, Jim Davidson, Jim Frew).
- Deprecation of support for the *jaquith* tape archive manager (Mike Olson, Jim Frew).

7.2. Data Manipulation

By 10 January 1993:

- Specification for a spatial transform package that handles conversion among different map projections and coordinate systems (Jim Frew).

7.3. Visualization

By 10 January 1993:

- Support for visualization of scientific data stored in POSTGRES using AVS (Peter Kochevar). Visualization will include rendering of surfaces and three-dimensional volumes, animation of time sequences, and display of data from model and satellite sources simultaneously.

8. Other Commitments

This section lists additional commitments that Sequoia Project members undertook during the development of this specification.

- Peter Kochevar will lead a seminar on how to use AVS for Sequoia Project members. This will be a ''bring-your-own-data'' affair; Peter will demonstrate how AVS can be used on data of immediate interest to researchers.
- Peter Kochevar will produce a specification for a Macintosh-like interface to a scientific visualization package like AVS.
- Jim Frew will produce a newsletter at arbitrary intervals describing the contents of the Bigfoot data store. Frew is responsible for allocation of space for scientific data.
- The sequoia-schema working group will come up with a standard external representation for large scientific data objects. Project members from SDSC, including Peter Kochevar and Zahid Ahmed, will participate. Interested researchers from other sites are encouraged to do so, as well.
- Scripps Institute of Oceanography will hire a staff person to handle visualization work there. Jim Frew and Richard Somerville will come up with a job description for this position.
- Scripps researchers would like to host an exchange student to work on visualization and scientific data management. Tim Barnett will find funding for this person. Jim Frew and Carl Staelin will attempt to find an interested student.
- Frank Davis will donate a copy of a report describing his work on wildfire mapping to the Project. Mike Olson will arrange for group access to an electronic version of this report.

  Other researchers are strongly encouraged to make copies of their publications available, as well.
- Peter Kochevar will work with the developers of AVS to remedy the problems Sequoia has with that software, particularly its memory demands.
- The computer science group at Berkeley has obtained an Alpha workstation for use by the Project. Performance numbers may not be published. All Sequoia Project members should have an account. Contact Jon Forrest to arrange for exclusive access to this machine if necessary.

  Jeff Dozier will run solar radiation models on the machine.
- Jim Frew will try to get an electronic glossary of Earth Science terminology from Hughes Aircraft for use by the computer scientists in the Project.
- Investigators at UCLA will determine what support they need to load and manage data on the Bigfoot server, and will arrange for that support with Berkeley.

- Peter Kochevar and Jim Frew will help Mike Stonebraker and Roberto Mechoso draft a proposal for an equipment grant from DEC to support Mechoso's NASA HPCC grant.
- Jim Frew, Ira Machefsky, and Peter Kochevar will arrange for DEC to contact Fabrice Cuq to get AVS support for the button- and dial-boxes that UCLA bought for their DECstation 5000s.
- Mike Olson will make available a binary-only distribution of the client and server software for POSTGRES, so that Sequoia sites need not support a full installation of the software to use Bigfoot.

Appendix A

Swat Team Trip Report

Earth Space Research Group
Santa Barbara, CA
Friday, 11 September 1992

1. Introduction

On Friday, 11 September 1992, the Sequoia 2000 Swat Team visited the Earth Space Research Group (ESRG) in Santa Barbara. The meeting included swat team members (Jim Frew, Mike Olson, Peter Kochevar, Duane Waliser, and Allison Woodruff) and members of ESRG and the Earth science community (Terry Figel, Bill Ohirok, Marty Landsfeld, Jim Davidson, Catherine Gautier, Didier Jourdan, and Mike Byers).

At the meeting, the Earth scientists presented their work. They emphasized the areas in which they need support from the computer science community. The area requiring the most support is the visualization of scientific data.

2. Background

At the last Sequoia retreat, researchers proposed several pilot projects to investigate global change using Sequoia hardware and software. These projects include

- understanding air/sea interactions and their effect on global climate;
- an analysis of the seasonal climatic cycle in the tropics, comparing model data to observational and remote sensing data;
- an analysis and model of net primary productivity, a measure of the biomass that exists at the bottom of the food chain; and
- mapping wildfires in California from satellite imagery.

An additional project has since been proposed by Warren White at Scripps Institute of Oceanography. This project would investigate the effects of El Niño on global climate systems. The El Niño project is most likely a second-year effort.

Only the air/sea interaction and tropical cycle projects were discussed in any detail on the 11 September meeting. The others will be covered at later swat team meetings.

We would like to see preliminary results for the four original pilot projects by the January 1993 Sequoia 2000 retreat. In order to get results by January, we must develop data visualization tools that support the pilot projects. Visualization means that data from various sources (satellite imagery, model output, observational data from ship tracks, and processed versions of all three) must be viewable simultaneously. ESRG scientists unanimously agree that they do not have the tools or expertise to do this kind of visualization right now.

3. ESRG Hardware and Organization

At the time of the meeting, ESRG owned five DECstation 5000/240s, four X terminals, and one DECstation 5000/25. The local network is Ethernet. They are connected to the Center for Remote Sensing and Environmental Optics (CRSEO) on the UC Santa Barbara campus by a T1 line. ESRG is located a few miles off-campus, and is unlikely to have a better link to the rest of

Sequoia anytime soon.

Catherine Gautier directs the ESRG research efforts. Frank Davis, Jeff Dozier, Dave Siegel, and Ray Smith are Sequoia investigators at the CRSEO lab on the UCSB campus.

4. The Pilot Projects

4.1. Air/Sea Interaction Project

The pilot project exploring air/sea interaction seeks to understand climate over open ocean. This project is of interest to several ESRG researchers. Three different research projects exploring different parts of air/sea interaction were presented.

4.1.1. Separating Snow and Ice from Cloud Cover using AVHRR

The Advanced Very High Resolution Radiometer (AVHRR) is a sensor on a weather satellite that collects images of the Earth in thermal and visual bands. Two satellites are normally flying at any time, both in polar orbits. A single AVHRR collects images of Earth surface temperature in a number of spectral bands. The satellite traces a curved swath over the surface of the globe, since it passes from pole to pole and the planet rotates beneath it. A single image spans the Earth's surface from horizon to horizon.

Viewed from space, clouds have a spectral and thermal signature that is very hard to distinguish from that of snow and ice on the ground. Catherine Gautier is trying to find a way to separate cloud cover from snow and ice in AVHRR images. In order to do this, she converts satellite images in different bands #a# and #b# to a *dimensionless parameter space*:

v~=~{{p sub a} ^-^ {p sub b}} over {p sub a}

The values #v# in parameter space have a spatial relationship to the pixels #p# in image space.

Catherine would like to be able to look at parameter and image space simultaneously. She would like to use color to highlight particular values in parameter space, and see the corresponding pixels highlighted in image space. This operation also applies in the other direction — highlighting parts of an image should cause the corresponding parameter values to be highlighted.

Catherine needs to be able to explore different parameter spaces and spectral bands easily. This means that the manipulation and visualization tools she uses must be interactive and relatively easy to use.

4.1.2. Monte Carlo Modelling of Photon Absorption

Bill Ohirok has a probabilistic model of atmospheric and Earth-surface photon absorption. He constructs a cube containing the volume of interest, and subdivides it into smaller cubes. Each of the smaller cubes has its own smaller coordinate system, and the coordinate systems of adjacent cubes may be different, so moving between them can cause a change in spatial resolution. Each subcube contains a mixture of aerosols (airborne particles), atmospheric gases, and water.

Bill shoots photons into the big cube. According to a probabilistic model, a photon bounces off a cube at a new angle or is absorbed by the cube. Some photons escape to space, some are absorbed by cubes in the model, and some are absorbed by the surface of the Earth. An absorbed photon contributes energy to the cube that absorbs it, affecting climate.

After a modelling run, Bill would like to visualize photon absorption and the trajectories of individual photons in three spatial dimensions and time. He would like to rotate the cube, fly through it, and incorporate results from separate runs of his model (for example, different wavelengths of light) in a single visualization.

At present, a single model run is for 100,000 photons. Eventually, Bill would like to do much larger (#10 sup 6# or bigger) modelling runs. He also plans to make the model more complex by making the Earth's surface non-uniform by introducing vegetation and soil cover.

The biggest problem Bill faces with respect to visualization are the size of his model output and the bandwidth required for interactive visualization of his results.

4.1.3. Analysis of Latent Heat Flux

Didier Jourdan is working on a project that requires high-speed database access from his climate modelling software. He would like to visualize both his source data and the results he produces simultaneously.

Didier would like to analyze *latent heat flux* over the ocean. *Heat flux* is the transfer of heat from one object to another; here, it refers to the transfer of heat between the Earth's surface and the atmosphere. *Latent* heat flux is heat transferred by the evaporation or condensation of water.

If you know the surface temperature of the sea, the wind speed, and the amount of water vapor in the air, you can compute the amount of evaporation or condensation that will occur. Didier's model uses sea surface temperature, wind speed, and water vapor content data to produce a map of latent heat flux over the ocean.

Didier would like to be able to select a point in his result and see all the input values that contributed to it. For example, he may want to answer the question, "Why does this output value look bogus?" This means that the modelling tool or database must understand the relationship between the source data and the model output. Also, Didier could ask for any of the ancestor data at any point, so the database should optimize its storage of ancestor data cleverly to make these lookups fast. For Didier's purposes, a minute is too long a time to wait for answers to questions of this sort.

4.2. Tropical Seasonal Cycle Project

Duane Waliser presented work in progress at UCLA on understanding the seasonal cycle in the tropics. This project uses data from different sources to predict or record seasonal climatic variations, and tries to understand the differences between the models and observations.

One data source is output from Roberto Mechoso's global climate model (GCM). GCM output is in three dimensions; the model records data at various elevations above given latitude and longitude positions. These data are regular (evenly spaced across the Earth's surface), span the entire globe, and have coarse spatial but fine temporal resolution.

Another data source is satellite imagery. Images have fine spatial resolution in two dimensions, and arbitrary temporal resolutions. In addition, pixels in satellite images may not be at regular intervals over the field of observation.

A third source of data is climatology observations from ship traffic. These data are irregular; observations in shipping lanes are more frequent, and hence more reliable, than observations at less well-travelled locations. In order to compensate for the irregularity of observations, monthly averages are computed and used as input to the tropical seasonal cycle project.

The main problem faced by researchers at UCLA with respect to visualization is the heterogeneity of their data. They have different spatial and temporal resolutions, and different data types, and they want to visualize them all simultaneously and compare individual values from each. In addition, they would like to do temporal animation on each data source.

5. Implications for Visualization

The pilot projects described in the last section all require good data visualization tools to succeed. Researchers must be able to visualize

- satellite images in multiple projections (for example, polar projections or equatorial projections spanning the prime meridian);
- geophysical products from processed satellite data (for example, Catherine's dimensionless parameter space from AVHRR images);
- observational data (shipboard temperature measurements); and
- model outputs (cloud and radiative heat transfer models).

These data sources may be very large, have varying temporal and spatial resolutions, and change resolution at internal boundaries. The scientists at ESRG would like to learn a single visualization tool that permits them to examine data from all of these sources, and to construct visualization images interactively.

5.1. Problems with Existing Packages

Terry Figel presented a summary of the problems with visualizing data from different satellite sources. Different satellites have different resolutions, both spatial and temporal. In addition, panoramic optical effects and surface curvature of the Earth distort images, particularly at the edges.

Existing packages for visualizing data have a number of problems, and no single package handles all data sources well. Among the problems with packages like IDL and TeraScan are

- projection singularities (strange behavior at the poles or prime meridian);
- excessive resource consumption — space or compute time — for examining large datasets;
- artificial limits of magnitude, such as being built for small-scale, rather than global, data visualization; and
- inflexible defaults, making it hard to build a visualization interactively.

5.2. Spatial Transformation on Images

Images of interest to Sequoia researchers are in different *geodetic projections*. Pixels in an image have a well-known spatial relationship to locations on the surface of the Earth. In some cases, pixels may be at known latitude and longitude. Pixel locations in any geodetic projection may always be converted to and from latitude and longitude.

When transforming an image from one geodetic projection to another, it may happen that either the source or the destination image contains holes. If an image is transformed into one with twice as fine a spatial resolution, the spaces between the original pixels must be filled in with sensible values. This problem is known as *resampling*.

Doing spatial transformations on images and resampling are well-understood. Sequoia researchers use images from many different sources, however, and no existing package handles transformations on all these sources. ESRG researchers need a generic spatial transform package that they can use to convert any image to and from some standard projection.

6. Action Items

Over the course of the day, the group agreed to pursue the following action items.

6.1. Specify a Generic Spatial Transform Package

The most important problem confronting ESRG scientists is the lack of a good tool for doing spatial transformation. By the January 1993 Sequoia retreat, we will write a document that describes

- existing software;
- its shortcomings; and
- whether we will write a generic spatial transform package ourselves.

In order to make this happen, everyone interested in spatial transformation will send their requirements to Jim Frew. Frew will generate a merged specification and search the network and literature to find out what exists now. Once the specification is ready, Mike Olson will find someone to write the software if we decide that we want to do that.

6.2. AVS Training

One powerful visualization tool with which we need more experience is AVS. This is an extensible system; it allows users to write modules that manipulate their data and to link these modules directly into AVS. Peter Kochevar suspects that software to handle many of the Sequoia visualization problems is already available for AVS.

Peter has volunteered to learn the system and to present a workshop on it before the January 1993 Sequoia retreat. Attendees will bring their own datasets with them, and Peter will show them how to use and extend AVS to manipulate their data.

6.3. Toolbox Interface for Visualization

ESRG scientists would like an interactive, Mac-like toolbox for constructing visualizations. For example, they would like to select from a palette of colors and icons for representing individual data values in a visualization.

Peter Kochevar will talk to the Earth scientists and will produce a specification for the toolbox interface. If possible, we will use an existing visualization package, like AVS, as the layer underneath this toolbox.

6.4. Bigfoot Dataset Newsletter

Jim Frew will produce a newsletter at arbitrary intervals describing the datasets that have been loaded on Bigfoot.

6.5. Data Loading

Duane Waliser and Terry Figel will load scientific project data onto the Bigfoot server in Berkeley. Researchers at ESRG will experiment with this arrangement to see how much more (or less) convenient it is than storing data on 9-track tapes locally. In addition, Peter Kochevar will use Duane's datasets to experiment with AVS.

7. Future Swat Team Meetings

The consensus was that the first swat team meeting was a success, and that the meetings should be continued. Visits to San Diego, UCLA, CRSEO in Santa Barbara, and Davis are planned for the near future.

Appendix B

Swat Team Trip Report

Scripps Institute of Oceanography
San Diego, CA
Monday, 29 September 1992 – Wednesday, 1 October 1992

1. Introduction

Beginning on Monday, 29 September 1992, the Sequoia 2000 Swat Team visited researchers at Scripps Institute of Oceanography (SIO) in San Diego. The meeting included swat team members (Jim Frew, Allison Woodruff, Duane Waliser, Peter Kochevar, and Mike Olson) as well as researchers from SIO and from the San Diego Supercomputer Center (SDSC).

The meeting spanned three days, with presentations by SIO researchers and computer scientists from other sitest on the state of Sequoia research. As was the case in Santa Barbara, visualization of scientific data received the most attention. However, emphasis was also placed on tools for analyzing data.

2. Background

Scripps is a department of the University of California at San Diego. Researchers there are largely independent of one another, but collaborate on some projects. Sequoia principal investigators at SIO are Richard Somerville, John Roads, Tim Barnett, and Warren White. Other researchers working on Sequoia-related projects include Dan Cayan, Shyh-Chin Chen, Nicholas Graham, Kyozo Ueyoshi, and Alejandro Pares-Sierra. There are a number of computer scientists at Scripps working on Sequoia-related projects. These include Mary Meyers, Tony Tubbs, Larry Riddle, Emelia Villaros-Bainto, Norm Hall, and Bill Borbridge.

Projects in progress are looking at land, oceanic, and atmospheric systems using observations and model data. Because of the size of the Scripps group and the diversity of the projects they are working on, this report will not describe individual projects in any detail. Instead, it will concentrate on areas in which the SIO researchers need common services.

3. Data and Storage in Bigfoot

The Bigfoot data store in Berkeley provides several terabytes of storage on several different devices, including magnetic disk and optical disk and tape jukeboxes. This storage is available to both database users and to users who want to store flat files outside of the database system. At present, the various devices have different interfaces; different programs are used to move data on and off of them. By January, Berkeley will provide FTP, NFS, and Postgres access to all supported devices.

Climate researchers should load the data they're interested in using. Frew will handle the allocation of space and will compile a list of all data available on the system. When data is loaded, it must include a description of its format. Other researchers will be free to use the data.

The computer scientists in Berkeley are very interested in moving large amounts of data into Bigfoot, so climate researchers are encouraged to move ahead with loading as quickly as they can. Since climatologists may not have the time or expertise to load data themselves, SIO will hire a staff member with DEC money to handle data loading.

Frew will compile the list of all data currently on Bigfoot and will give it to Warren White and other interested researchers immediately.

4. Datatypes and Schema Issues

To make it easier to share data, researchers should convert their datasets into a standard format that all Sequoia applications can understand. In order to do this, the project needs to define a set of datatypes (such as *temperature*, *depth*, and *salinity*) and their representations (including number of bits stored, precision, and encoding of units). A group of Sequoia members, chaired by Frew, is working on designing these types. The group is called *sequoia-schema*, and interested parties are encouraged to join.

The schema group has so far designed common table formats for data stored in Postgres tables. Work on datatypes and representations is proceeding. Until this work is finished, researchers should load data using whatever type is currently available, and makes sense for the values stored. For example, latitude and longitude might be stored as four-byte integer values.

If you have questions when you choose types for data stored in Postgres, contact Mike Olson or Frew for help.

John Roads pointed out that common standard formats won't be useful in all cases. For example, to checkpoint a model run, it may be necessary to store temperatures as double-precision floating point values, since the model's internal state must be restored from the checkpoint exactly. In cases like this, researchers should choose the datatype required to do the job. The goal of standardization is to make it easy to share data, not to make it impossible to do work.

5. Large Objects

Since many of the datasets Sequoia researchers work with are large — for example, grid output from Global Climate Model runs — the Sequoia standard schema should include standards for large object data. Large objects may be stored inside Postgres or as flat Unix files.

There are several issues in representing large objects:

- We would like all large objects of a single variety to have the same representation inside the database system.
- We would like all large objects of a single variety to have the same representation outside of the database system. That is, we should support a common import/export format for large object data, so that it can be shared with researchers who are not on Project Sequoia.
- We would like to avoid designing entirely new representations for large object data.

Sequoia should permit scientists to work with data in familiar formats. At the same time, to encourage sharing of data among different projects, it is important that a standard interchange format be developed. In principle, any dataset in the Sequoia database should be available in a standard format, and may also be available in other formats.

The sequoia-schema working group will come up with a standard representation for large object data. Interested researchers are encouraged to help define this standard.

Two important issues in the representation of large objects are *metadata management*, or how information describing the large object data is stored, and *representation*, or appropriate formats for the actual large object data. Much work has been done elsewhere on managing self-describing data objects, which include a description of the data format with the data itself. Net-CDF and HDF are two standards for representing self-describing objects. Zahid Ahmed and Peter Kochevar of SDSC have some experience in this area, and will work with the Sequoia schema group to incorporate large object representations into the standard schema.

Large object types of interest to Sequoia researchers include raster images and volumetric data. Volumetric data is a *k*-dimensional grid of *n*-tuples — for example, output from a GCM run, which is a four-dimensional volume with wind velocity, temperature, and other measurements at every point in the grid. Volumetric data is of especial interest. Mike Olson will look at providing a general-purpose gridded large object data type by the end of October. If none can be written at Berkeley by then, the computer scientists at SDSC may write one.

6. Scientific Data Management

Sequoia researchers at SIO all do preprocessing and conversion of datasets they receive, visualize them, and would like to run statistical analyses on them after they identify interesting features in the visualization. This section describes the processing steps common to all SIO researchers and lists the services that the Sequoia computer science community should provide.

6.1. Data Conversion

Data from external sources must typically be processed in order to be useful to climate researchers. The first processing step requires translating the data from the format in which it arrives to a format that can be managed by the hardware and software used for analysis. Important issues include the size and precision of integer and floating point values, how to store dates and times, and minimizing the size of the processed dataset.

In general, this processing will happen before the data is loaded into Bigfoot, and the storage system will not have to provide conversion routines. Researchers who want the storage system to handle conversion automatically must provide routines to do the conversion; these can be invoked automatically when new data arrives.

Data conversion is not a very hard problem. It is related to the Sequoia standard schema, since the schema will define the format of the output dataset. However, no serious work is required here.

6.2. Data Transformation and Interpolation Techniques

Once the data is in a format suitable for analysis by the software tools that a researcher uses, it must generally be transformed in some way. The transformations of interest to Sequoia researchers at SIO are the same as those used by researchers at the Earth Space Research Group in Santa Barbara.

Data that arrives on an irregular grid must be mapped to a regular grid in some way. Since the two grids have different spatial resolutions, actual data values must be interpolated to produce values on the new grid.

The interpolation technique used by Warren White and other researchers at Scripps is called *optimal interpolation*. First, a variogram is produced from the data that represents how similar values are at different distances from one another. By finding the sill of the variogram[2], a researcher can estimate the degree to which values are affected by their neighbors, and the distance at which such interaction stops.

Second, the variogram is used to produce a weighting curve that assigns more weight to nearby points than to distant ones[3].

Third, the weighting curve is used on points on the input grid to produce points on the output grid. Since the weighting curve is bell-shaped and typically falls off quickly, the number of source points required to compute destination points is small.

This interpolation technique yields much better results than less sophisticated techniques like bilinear interpolation. SIO researchers have code that does optimal interpolation, and will contribute it to Sequoia for inclusion in a data processing tool.

Optimal interpolation produces a good weighting curve to use in computing target values. Other interpolation techniques use other curves, but typically apply them to source points in the same way. Sequoia should develop a general-purpose interpolation tool for use on scientific data. Users should be able to select the variety of curve they want (for example, optimal or bilinear) and produce gridded output data easily. Frew will work with Warren White at SIO to specify the interface for this tool and to begin constructing it.

6.3. Visualization

This section describes current visualization efforts at SIO. Section 7 lists nine visualization benchmarks proposed for Sequoia by Warren White.

Data visualization is a tool to help scientists identify interesting properties that they can then analyze more rigorously. SIO scientists feel strongly that any visualization tool must be closely integrated with the analysis tools described in the next section. Peter Kochevar pointed out that visualizing data and visualizing the results of data analysis are equally feasible, and should both be supported.

SIO researchers are using many different visualization tools, including AVS, IDL, NCAR graphics, and other packages developed by researchers elsewhere or at Scripps. There is a good deal of IDL expertise. Users of NCAR graphics like it because of its strong bias toward the Earth sciences. Of the packages in common use, AVS and IDL are by far the easiest to extend, but there is not much experience with AVS so far.

Peter Kochevar of SDSC is working with John Roads of SIO to do visualization of some of John's regional climatological data. Once this project is completed, Peter will give a seminar at Scripps on how AVS can be used on data from other researchers.

In order to make it easier for Scripps to use to AVS, they will hire someone to become the local visualization expert. That person will provide support for the scientists, and will work with computer scientists at other sites. Jim Frew will work with Richard Somerville to generate a job description for this position.

In the short term, Scripps researchers are interested in hosting an exchange student from some other site to help with scientific data management and visualization. Tim Barnett has agreed to find funding for this person. Mike Olson and Jim Frew will to find an interested

[2] In computer science terms, the knee of the curve

[3] Of course, the variogram need not depend on distance, but it generally does.

student[4], and Barnett will work with them to be sure that the student is right for the position.

6.4. Statistical Analysis of Data

Looking at data with a visualization tool suggests features that can be more rigorously analyzed. This section describes some of the types of analysis that SIO researchers do.

First, a common analysis technique is to concentrate on anomalies, rather than actual measurements. Anomalies are differences from mean (or otherwise expected) values. Anomalies tend to cluster, and anomalies in one domain (sea surface temperature, for example) tend to be clustered with anomalies in other domains (such as precipitation).

Second, virtually all SIO researchers use Complex Empirical Orthogonal Functions (CEOF) to simplify and characterize the data they work with. This technique is similar to Fourier series analysis in computer science, except that CEOF may choose to use non-sinusoidal curves in constructing the target curve. For a given dataset, values are plotted on some grid, and the curve that passes through the points is constructed. Next, the curve's principal components are extracted. Generally, the first few components provide more than ninety percent of the curve's shape. This small CEOF set can then be used to characterize the dataset in subsequent analysis and in models.

Finally, several different statistical and mathematical techniques are used frequently by SIO scientists. Correlations among datasets, analysis of variance (ANOVA), and function integration and differentiation are common analytical techniques.

All of these analytical techniques should be part of the data management toolkit.

6.5. Summary

Using data in science involves a number of different processing steps. Any scientific data management tool adopted by Sequoia should provide services for each step. At SIO, researchers commonly must convert data to manageable types, transform it to a reference system with which they can work, visualize the data, and then analyze it. Current systems make preprocessing, visualization, and analysis entirely distinct. Sequoia should try to integrate these steps, so that they could all be done using a single toolkit.

7. Nine Visualization Benchmarks

Warren White produced a list of nine visualization benchmarks for Sequoia to run in the short term. These benchmarks are of strong interest to researchers at SIO, and are likely to be interesting to researchers at other sites, as well. Researchers at other sites are strongly encouraged to elaborate on Warren's list, or to draw up a similar list of their own.

This section briefly describes the nine benchmarks for the computer science community. Climatologists already familiar with the discipline's terminology may find Warren's original document, in the next appendix, more interesting.

7.1. Maintain a Mean Annual Cycle Dataset

For a given dataset that represents some mean annual cycle (*e.g.*, equatorial rain measurements), researchers want to be able to acquire new measurements, visualize and analyze them, and incorporate them into the mean annual cycle dataset.

---

[4] So if you're interested, send us mail. San Diego winters are mild.

This requires that the new measurements be projected on some map. Differences from the mean annual cycle are produced, and these anomalies are then re-gridded, using optimal interpolation, so that they are at the same resolution as the mean annual cycle data. The anomalies can now be examined with a visualization tool.

If visualization turns up questionable values, the original measurements that produced them should be annotated by the researcher, and the questionable values discarded. Optimal interpolation and visualization repeat until the dataset is acceptable.

At this point, the processed new measurements can be added into the mean annual cycle dataset, producing a more accurate version of the cycle.

Jim Frew reports that Berkeley is working on handling the map projection problem. Warren volunteered to donate his optimal interpolation code to the project.

7.2. 4D Display

This benchmark requires that three-dimensional data (such as ocean bathymetry) be plotted over the globe according to some map projection. After the ocean floor is rendered, three-dimensional shapes must be drawn in the water. For example, temperature variation can be displayed geometrically. All the points at which water temperature is 6°C in a given region form an *isosurface*, and this isosurface can be drawn.

It should also be possible to animate these pictures, to watch the isosurface move over time. At any timestep, the user should be able to interrupt the animation, and slice through the displayed image to see a planar plot of values.

From this visualization, researchers will want to compute statistical properties of the data and the CEOF. The visualization tool must support sophisticated analysis.

The project should acquire high-resolution maps of political boundaries, geographical boundaries, and land topography and ocean bathymetry for the whole planet.

7.3. Surface-Volume Animations

This benchmark requires that volume and surface data (for example, the ocean surface and the atmosphere above it) be visualized simultaneously. Data from the two sources must be mapped to a common grid and displayed. Changes in time must then be shown in step.

An example is the relationship between heat flux at the sea surface and wind velocity and humidity in the atmosphere. In slow-moving, wet air, less sea water evaporates, minimizing heat loss at the ocean's surface.

Again, statistical analysis follows naturally from the visualization.

7.4. Volume-Volume Animations

This benchmark extends the last one. Just as the atmosphere and ocean surface are related systems, the atmosphere and ocean volume are. The same sorts of animations described in the previous section should be possible on two three-dimensional datasets.

7.5. Large Numbers of Datasets

This benchmark requires that data from #n^<^12# different sources be converted to the same grid, and that statistical analyses on them be done (for example, are different pairs strongly correlated? At what space or time scale?).

Visualizing this many datasets simultaneously is very hard; so far, we have no good way of dealing with clutter. According to Warren, it would be sufficient (and an improvement over what's currently possible) simply to tile the screen into #n# regions, and display one dataset in each region. Animations on the datasets should happen in lockstep.

CEOF analysis on the #n# datasets should be supported.

7.6. Model/Data Combination

This benchmark is similar to one proposed at the swat team's trip to Santa Barbara. Model data and observations must be used simultaneously in a single analysis and visualization. This requires gridding data from different sources in the same way, and displaying several different datasets simultaneously.

7.7. Model/Data Assimilation

This is an interactive process by which a researcher can refine a climatology model. First, the model is run to produce some data. Then real observations are collected. The model output and actual measurements are mapped to the same grid, and the differences between model output and observed values are displayed. From these differences, the researcher changes formulae in the model to produce answers more like the observed values.

This process can be repeated many times.

7.8. Climate Model Forecasting

This benchmark uses the previous one to refine a model, and then runs the model into the future. The values output from the model are a forecast. The correlation between the forecast and observed data can be computed. This technique can be used to do "hindcasting," in which a model produces a forecast for some year in the past, and the model's output is compared to measurements collected that year.

7.9. Designing Observation Networks

An important issue in the sciences is determining where, and how often, observations should be made. Temperatures vary little over small distances or time periods, so measuring temperature every inch or every second is not productive. Using analysis and visualization, the best scale for collecting measurements can be determined.

Using statistical analysis techniques, the space/time decorrelation scale for a given dataset can be computed. This scale can be displayed by the visualization tool. The researcher can identify the point on the curve at which changes become important, and can use that information to design a network of observing stations.

8. Miscellaneous Issues

8.1. Text

Richard Somerville expressed interest in how Sequoia will support textual data. Allison Woodruff described her research plans for him, which include an intelligent indexing technique for textual data. Richard is interested in storing text in Bigfoot, and using a smart bibliography system to construct customized textbooks for classes.

8.2. Documentation

SIO researchers encouraged Berkeley to produce a document describing the general philosophy and lore of the project. Such a document would include a guide to the available software, why it was chosen, and what to do about it if you don't like it. Mike Olson agreed to contribute a high-level guide to Postgres to such a document.

Jim Frew agreed to supply a list of all data currently available on Bigfoot, and to produce regular updates as new data is loaded. The anonymous FTP directories currently contain some of this information, and should eventually become a project document.

8.3. Computer Systems

Jim Frew controls hardware and storage space for use by project investigators. He will come up with a formal guide to both, including some information on where programs should be run and how data storage is allocated.

9. Open Issues

The day after the swat team trip to San Diego ended (October 1), a meeting of the visualization group was held via PictureTel. At the meeting, the swat team was charged with investigating support by the visualization system for data lineage and computational steering. Under close examination, these two terms turned out not to have definitions.

At the upcoming swat team meetings in Los Angeles, Santa Barbara, and Davis, definitions for these terms should be agreed upon, and a proposal for their support by the visualization package should be made.

Appendix C

Visualization Benchmarks for Global and Regional Climate Change

This section contains the nine visualization benchmarks as they were proposed by Warren White. The benchmarks list specific applications and the names of researchers who would benefit from the benchmark. The researcher who can give more details on the benchmark is named in **bold face**.

C.1. Map Non-Uniformly Distributed Data

Display data locations, interpolate data, form anomalies, contour data, evaluate data, re-interpolate and re-contour, combine anomalies and mean.

Specific applications: map ocean physical, chemical, and biological station data onto a uniform grid at a uniform time; place time series data at ocean buoy stations or airports or radiosond stations onto a uniform grid at a uniform time; place satellite data onto a uniform grid at a uniform time.

Interested researchers: **Norm Hall**, Warren White, John Roads, Dan Cayan.

C.1.a. Take any 4D data set distributed randomly over some specific domain and display the data locations on some projection (be able to post the values if desired).

C.1.b. Calculate the mean, variance, and covariance statistics (*i.e.*, spatial and temporal decorrelation scales and the signal-to-noise ratios).

C.1.c. Form anomalies from the mean annual cycle.

C.1.d. Interpolate anomaly data onto a specific uniform 4D grid using optimal interpolation, and compute the interpolation errors.

C.1.e. Contour or colorize the gridded anomaly values and their errors (be able to overlay with original data locations).

C.1.f. Open a data window on individual data locations to check questionable anomaly values, and be able to insert a quality control flag (metadata) on the original data.

C.1.g. Re-interpolate and re-contour anomaly data of known quality (*i.e.*, repeat steps c to f), producing quality-controlled 4D gridded fields of anomaly data.

C.1.h. Add the anomalies back onto the mean annual cycle.

C.2. 4D Display

Topographic relief, render 3D volume, animate 3D volume, slice 4D volume, correlation matrix.

Specific applications: render salinity tongues in the ocean; render current jets in the ocean and atmosphere; render temperature anomalies in the ocean and atmosphere.

Interested researchers: **John Roads**, Shyh-Chin Chen, Tim Barnett, Richard Somerville, Norm Hall, Warren White, Kyozo Ueyoshi, Alejandro Pares-Sierra.

C.2.a. Plot the surface of the ocean or land bathymetry as flat boundaries or in relief over any portion (or all) of the globe on #n# map projections (*e.g.*, spherical, Mercator, Lambert equal-area).

C.2.b. Render the surface of a scalar contour in 3 spatial dimensions encompassing a volume, allowing resulting “worms” and “slugs” to be traced through the 3D space of the ocean or atmosphere system.

C.2.c. Be able to enter into the 3D space to examine rendered scalar volumes from a variety of different angles and zoomed distances.

C.2.d. If data have a time dimension (*i.e.*, the 4th dimension), allow the “worms” and “slugs” to “burrow” (extending the metaphor) through the 4D space of the ocean or atmosphere system.

C.2.e. Display a contour of a 2D slice through the 4D space-time volume.

C.2.f. Compute the auto-covariance matrix, and display the complex EOF.

C.3. Surface Effect to Volume Response

Surface contour, render 3D volume, join fields, animate join fields, cross-correlation of surface with volume.

Specific applications: Map SST anomalies over the global ocean, with air temperature anomalies rendered over the entire globe; map wind stress curl over the North Pacific Ocean and render the subsurface temperature anomalies below it; map snow cover anomalies over the Himalayas and render the air temperature anomalies above the entire globe.

Interested researchers: **Richard Somerville**, Nicholas Graham, John Roads, Tim Barnet, Shyh-Chin Chen, Kyozo Ueyoshi, Alejandro Pares-Sierra.

C.3.a. Plot the contour or relief of surface effects (*e.g.*, SST, radiation flux, primary productivity, snow cover, ice cover, biomass, population density, etc.).

C.3.b. Render the 3D volume response, either below the surface in the ocean or above the surface in the atmosphere.

C.3.c. Insert the contoured surface and the rendered volume on the same window.

C.3.d. Animate the surface and the 3D volume above or below it.

C.3.e. Compute the cross-covariance matrix between the surface and the volume either above or below it, and display the multi-variate CEOF.

C.4. Ocean Volume and Atmosphere Volume Interaction

Render ocean 3D volume, render atmosphere 3D volume, join the two fields, animate ocean-atmosphere interaction, cross-correlation between ocean and atmosphere.

Specific applications: Animate subsurface temperature anomalies in the global ocean and air temperature (or clouds, moisture content, etc.) anomalies in the global atmosphere above it to see how anomalies in one fluid medium influence those in the other.

Interested researchers: **Nicholas Graham**, Tim Barnett, Warren White, John Roads, Kyozo Ueysohi, Alejandro Pares-Sierra.

C.4.a. Render 3D volume of some oceanic anomalies over some portion (or all) of the global ocean on one of #n# projections.

C.4.b. Render 3D volume of some atmospheric anomalies over some portion (or all) of the global ocean on the same projection.

C.4.c. Merge the two rendered volume anomalies over the same or different portions of the globe.

C.4.d. Animate the rendered volumes in both fluid media.

C.4.e. Compute the cross-covariance matrix between oceanic and atmospheric anomalies, and display the multivariate CEOF.

C.5. Multiple Interactions

Define #n# 2D datasets, normalize #n# 2D datasets, cross-correlation statistics, multivariate principal components.

Specific applications: Examine the interaction between ocean temperature anomalies off the coast of California, wind-air temperature/cloud-precipitation anomalies over California, snow cover and biomass anomalies on the land of California, and river runoff and reservoir capacity anomalies.

Interested researchers: **Dan Cayan**, Tim Barnett, Nicholas Graham, Warren White, John Roads.

C.5.a. Define #n^<^12# datasets of gridded 2D data over some portion (or all) of the globe.

C.5.b. Normalize the datasets and display ANOVA statistics (including decorrelation space and time scales) in a table with #n# entries.

C.5.c. Compute the cross-covariance matrix between the #n# datasets, and display the possible interactions (*i.e.*, maximum cross-correlation at #x# space lag and #y# time lag) in an #n ^times^ n# matrix.

C.5.d. Compute the multivariate CEOF between the #n# datasets, and display the principal components in a table with #2n# spatial patterns and one time sequence.

C.6. Model-Data Intercomparison

Define model and observed datasets, difference model and observed datasets, cross-correlation statistics of differences, multivariate principal components of differences.

Specific applications: examine the ability of an ocean model to simulate ocean temperature anomalies off the coast of California measured from AVHRR; examine the ability of an ocean-atmosphere coupled model to simulate the observed upper ocean temperature and lower atmosphere winds over the global equatorial wave.

Interested researchers: **Tim Barnett**, Nicholas Graham, Alejandro Pares-Sierra, John Roads, Kyozo Ueyoshi.

C.6.a. Define model and observed datasets in 4D over some portion (or all) of the globe.

C.6.b. Difference the model and observed datasets and display ANOVA statistics (including decorrelation space and time scales).

C.6.c. Compute the cross-covariance matrix between the model and observed datasets, and display possible interactions (*i.e.*, maximum cross-correlation at #x# space lag and #y# time lag).

C.6.d. Compute the multivariate CEOF between the model and observed datasets, and display the principal components in a table with #2n# spatial patterns and one time sequence.

C.7. Model-Data Assimilation

Operate prognostic climate model, define model and observed datasets, interpolate model data to location of observed data, difference model and observed data, decorrelation statistics of differences, interpolate differences to model grid, update model first-guess field.

Specific applications: Conduct model-data assimilation on a global ocean model using observed XBT temperature data collected over the ocean by the Global Ocean Observing System (GOOS).

Interested researchers: **Warren White**, Tim Barnett, Nicholas Graham, John Roads.

C.7.a. Load into memory a community prognostic climate (coupled) model, integrate the model, and display #n# fields.

C.7.b. Define model datasets that will be updated by observed datasets.

C.7.c. Interpolate the model data to the location of non-uniformly distributed observed data.

C.7.d. Difference the interpolated model data with the observed data, and display decorrelation statistics (including the signal-to-noise ratio).

C.7.e. Interpolate the model-data differences onto the model grid using optimal interpolation.

C.7.f. Update the first-guess field of the model with model-data differences under the nudging criterion.

C.7.g. Integrate the model over the time period coincident with the observed data.

C.8. Climate Model Forecasting

Operate prognostic climate model, assimilate observational datasets, integrate the model into the future, determine skill of forecast.

Specific applications: Conduct hindcasting over the global ocean-atmosphere system for the period 1972 to 1992, using a coupled model assimilated with subsurface temperature and surface wind stress.

Interested researchers: **Tim Barnett**, Nicholas Graham, Warren White.

C.8.a. Load into memory a community prognostic climate (coupled) model, integrate the model, and display #n# fields.

C.8.b. Call the model-data assimilation recipe, yielding C.7.e above.

C.8.c. Integrate the coupled model from model-data assimilated state into the future in the absence of observed data, and display #n# forecast fields at #m# lag times.

C.8.d. Compute the cross-correlation of observed data with the forecasted fields.

C.9. Observing System Network Design

Decorrelation scales, dynamical decorrelation scales, assimilate observational datasets, integrate the model into the future, determine skill of forecast.

Specific applications: Design an upper ocean temperature observing system over the global ocean in support of the Global Ocean Observing System; design a wind vector measuring buoy array along the coast of California in support of the Coastal Ocean Program.

Interested researchers: **Warren White**, John Roads, Alejandro Pares-Sierra.

C.9.a. Examine fields of decorrelation time-space scales and signal-to-noise ratios of differences from A.7.d above (or from the anomalies in A.1.f above).

C.9.b. Calculate the interpolation error as a function of data density, given the decorrelation scales and signal-to-noise ratios, for every location where the decorrelation scales are given.

Appendix D

Swat Team Trip Report

Center for Remote Sensing and Environmental Optics
Santa Barbara, CA
Friday, 30 October 1992

1. Introduction

On Friday, 30 October 1992, the Sequoia 2000 Swat Team visited the Center for Remote Sensing and Environmental Optics (CRSEO) in Santa Barbara. The meeting included swat team members (Jim Frew, Mike Olson, Peter Kochevar, Duane Waliser, and Allison Woodruff) and researchers and staff members at CRSEO (Jim Davidson, Frank Davis, Jeff Dozier, Steve Miley, Walter Rosenthal, Jiancheng (J.C.) Shi, and Dave Siegel).

CRSEO researchers presented their research and discussed areas in which they need more computer support. As was the case at previous meetings, visualization was discussed at some length. At this meeting, two new topics were also discussed with the swat team — computational steering and data lineage. Definitions of these terms and some early ideas for supporting them appear later in this report. Researchers at other sites, who have not yet had a chance to discuss these topics, are encouraged to send their opinions to Mike Olson (mike_olson@postgres.CS.Berkeley.EDU) for inclusion in the final swat team benchmark specification.

2. Overview of CRSEO

The Center for Remote Sensing and Environmental Optics is an independent Organized Research Unit (ORU) on the UCSB campus. Researchers at CRSEO, like those at Scripps, are independent of one another, but sometimes collaborate on projects and in general try to exploit similarites among their specialties. For example, mathematical modelling techniques developed for one specialty are often directly useful to researchers in other specialties.

3. Data Lineage

At the 1 October meeting of the Sequoia visualization group, the swat team was charged with defining the terms "data lineage" and "computational steering," and with incorporating them into the benchmark specification. In order to get as much input as possible, these topics were covered in the morning meeting at CRSEO, with the entire group.

Keeping track of data lineage means keeping track of metadata that describe data of interest to a researcher. This metadata should be tightly bound to the data it describes, so that it is easy to examine, and so that it follows the data on its path through the system.

Some examples of metadata that must be maintained include:

- A description of the data's format and of any devices used to collect it.
- The processing history of the data, including the code that was executed to transform the data. Note that this description can be arbitrarily fine-grained. For example, it should be possible to determine, for a single pixel of a processed image, what input image pixels contributed to its value, and what algorithm was applied to compute the pixel's value.

- Annotations assigned by humans to data. For example, if a set of field measurements were collected by someone unreliable, the data could be tagged as suspect. Similarly, if an image is oversampled to very high resolution, individual pixel values are untrustworthy, and should be marked as such.

For data stored in POSTGRES, complete time history is stored, but the connections between each step in time are not. In order to maintain the processing history of data, POSTGRES for Sequoia must be extended to maintain these connections automatically. In addition, code that is used to process data should be stored in the database system, so that the correct version of it can be located when processing history is being explored.

Annotation is inherently an ad-hoc operation; some user of data simply wants to attach descriptive text to some or all of it. An appropriate metaphor is the use of Post-It notes. As data is transformed by processing, all of the Post-It notes attached to it should follow it. For example, if a user determines that raw data is entirely unreliable, then any data derived from it will be equally unreliable, and should be so tagged.

An important issue, not resolved at the swat team meeting, is how the enormous amount of metadata for any datum can be managed by a user. Some interface to data lineage must be provided. Designers of Sequoia interface tools should consider this issue as they construct their systems.

4. Computational Steering

Steering a computation means that its course can be changed in real time, interactively, by a user.

Consider a global climate model. As they currently run, these models start from some initial state and run for some number of timesteps. The user cannot interactively change *variables* in the model (*e.g.*, the salinity of a particular ocean cell) or *parameters* of the model (*e.g.*, the solar constant[5]). A model changes the values of its variables, but not of its parameters.

This means that a researcher must re-run the model from scratch, often modifying a computer program to change the parameter or variable of interest. If variables and parameters were under interactive control, then models could be more easily used to explore system behavior.

A number of interesting observations were made at this point. First, viewing models in this way makes them appear very much like spreadsheets. Values in individual cells are linked by systems of equations. Spreadsheet programs permit users to change either the equations (in model parlance, the parameters) or the cell contents (variables). This seems to be a powerful metaphor, and we should consider studying spreadsheets in the context of Sequoia.

Second, Peter Kochevar pointed out that cellular automata are another example of grid cell based computation. In a simple system of automata, a checkerboard is populated with some number of citizens, with no more than one per square. At each time step, new citizens are born on empty squares or old citizens die, depending on the population in the immediate neighborhood. State transitions like these are similar to those in spreadsheets. Cellular automata may be another interesting abstraction to explore for Sequoia.

No actual plan for attacking computational steering was proposed. In the short term — that is, by the January retreat — we do not plan to rewrite existing global climate models as

[5] Climatologists find it useful to change the solar constant, but nevertheless insist on calling it that.

spreadsheets. Instead, project members should continue to think about new modelling paradigms. In addition, new data manipulation software should expose interesting knobs and switches to the user. AVS modules should allow dynamic configuration, to the extent possible.

5. Visualization

CRSEO researchers need the same kinds of visualization support that are required by their colleagues at the other Sequoia sites. The requirements include:

- A tool to convert observational and model data to and from different projections. This includes interpolation of images. As was the case at previous swat team meetings, Jim Frew agreed to oversee the implementation of such a tool. A student at Berkeley has begun working on this. The work will be based on GCTP, a transform package written in FORTRAN and available for anonymous ftp from postgres.CS.Berkeley.EDU in directory frew/src, file gctp.tar.Z.
- The ability to display objects in a format other than that in which they are stored. As a simple example, double-precision floating point values should be mapped to eight-bit colormap indices for rendering under X11.
- The ability to visualize multi-dimensional data easily, using techniques like volume surface rendering, slicing, and animation over time.

Tools currently in use at CRSEO include

- IPW, a home-brew package that uses Unix pipelines to compose complex visualizations from simple tools;
- IDL, the interpreted language in use at other sites;
- AVS, the visual dataflow system;
- spreadsheets; and
- POSTGRES.

Jim Davidson recently gave a seminar on the similarities and differences between AVS and IPW. Most of the expertise at CRSEO is with IPW, and there is strong interest in turning some IPW modules into AVS modules. The current version of AVS consumes far too much memory, however, so this may not be a good idea in the very short term.

Peter Kochevar recommends two strategies for resolving memory problems under AVS. First, sites should load one or two machines with a lot of memory (128 megabytes makes it possible to use AVS; 64 megabytes is not enough) and use these to do AVS visualization. Second, AVS is working on the next release of their system, and Sequoia has close contacts with the company. We should state explicitly to the developers that memory consumption is a serious problem in the current release.

6. Projects at CRSEO

This section summarizes the projects presented to the swat team during our visit.

6.1. Net Primary Productivity and Wildfire Detection

Frank Davis is using AVHRR satellite images to compute a "vegetation index." AVHRR images are released as bi-weekly composites, meaning that images from two weeks are combined to produce a single cloud-free image. Davis uses two image bands to compute the greenness of a given pixel.

In boreal (northern) forests, the resulting vegetative index maps can be used to identify places that have been burned by wildfires, but this is because wildfires in boreal forests are large and reduce dense forest to ash. In forested regions of the California Sierras, fires are smaller and occur in mixed terrain.

Davis' project is trying to compute the footprint of a wildfire from vegetative index maps for the Sierras. In order to make this easier, actual wildfires are identified by looking for areas saturated in the thermal infrared band. When the fire has burned out or been extinguished, algorithms can concentrate on the region around the fire to identify its borders. Another strategy is to look for wildfires only where they are likely to happen. Crop harvest in agricultural regions looks remarkably similar to wildfire from the satellite images, but wildfires in these areas are improbable.

In order to visualize the resulting wildfire data, Davis would like to animate the vegetative index maps over time. He would like to be able to zoom in and out on the animation interactively.

Davis has a report that will be published in a book by Springer-Verlag describing this work. Mike Olson has agreed to get an electronic copy of this report so that interested parties can get it by FTP from postgres.

6.2. Modelling Penetration of Radiation into the Ocean

Dave Siegel is modelling the penetration of solar radiation into upper layers of the ocean. Most global climate models currently ignore differences in absorption and reflectance at different ocean depths. Siegel is modeling solar radiation penetration using Catherine Gautier's surface radiation flux data, combined with Coastal Zone Color Scanner (CZCS) chlorophyll imagery and ocean mixed layer depth information. The immediate goal of this project is to evaluate the amount of heat that penetrates through the mixed layer and so does not participate in air-sea interaction processes. In many oceanographic regions (in particular, the tropics), these fluxes can be quite large (#O (50 ^ W {m sup {-2}})#). The project's ultimate goal is to develop the means to assimilate data from future ocean color imagers, such as the SeaWiFS (Sea Viewing Wide Field of View Sensor), into general circulation models.

Siegel's primary Sequoia-related interest is in the combination of three different kinds of data for his model.

6.3. Turbulence Modelling

Dave Siegel is also modelling and visualizing oceanic turbulence in a triply periodic cube. The goal of this modelling effort is to mechanistically predict vertical mixing coefficients of ocean circulation models. Most available fluid dynamics packages model airflow over wings, and are not useful for this sort of modelling. The model consumes large amounts of Cray time, making it hard to do interactive analysis or computational steering. When Siegel wants to explore details of the model more carefully, he must change the model and restart it on the Cray. Interesting analyses must be run on the supercomputer because the saved model state is not sufficient for local analysis.

Frew notes that if complete model state were stored in the database, then subsets of interest could be extracted easily using selection criteria. This kind of support would make new types of analysis possible; for example, Siegel currently examines only single frames at a time, but if all frames were available, he could look for standing waves in the system.

6.4. Estimating Snow Cover from Satellite Images

Walter Rosenthal is analyzing Thematic Mapper satellite images to compute the location and water content of snow. His approach is to build a simple decision tree that identifies snow and its water content, and then to use that tree to analyze new satellite images. A single image is analyzed numerically using a divide-and-conquer algorithm that runs very slowly, but groups pixels that are most similar into the same subtree. The result is that given a coarse description of a new pixel, the pixel can be located in the correct subtree.

Walter has succeeded in building such a tree from TM satellite data. The next step is to use the tree to analyze other TM images, and to figure out whether the snow location and water content guesses that the tree makes are correct. At the time of the swat team meeting, he had not yet done this, but expects to do so in the very near future.

Walter would like storage and analysis support from Sequoia for this process. Storage is not a problem. Walter uses the S statistical package to construct the decision tree, and S is not integrated into either POSTGRES or the set of Sequoia tools available for use by researchers. We should provide statistical processing using S.

6.5. Estimating Snow Water Content from Radar Images

J.C. Shi is using images from Synthetic Aperture Radar (SAR) to find snow and to compute water content. This collection device allows users of the data to extract any pair of transmitted and received polarizations of the image. Varying the polarization during analysis allows Shi to classify snow, ice, and rock without having topographic data. In addition, snow water content can be computed.

SAR images are in a format similar to Thematic Mapper images, but must be managed differently. The inclusion of polarization information, and the use of the Stokes matrix to allow any intermediate polarization to be extracted, make this kind of data inherently not subsamplable.

7. Miscellaneous

This section covers minor issues that were discussed during the day.

7.1. Sequoia Storage Issues

Sites that want to load data onto Bigfoot right now should use either Jaquith (to store data on the Metrum jukebox) or icopy (to store data under POSTGRES). NFS support for the Inversion file system on top of POSTGRES will be available by the end of the year.

Steve Miley would like to understand policy on the use of Bigfoot. For example, should Bigfoot be used in real time by applications, or should data be fetched from Bigfoot and cached locally?

Both Jim Davidson and Steve Miley encouraged the storage group to come up with a strategy for staging data at Sequoia sites that does not require management by the user. For example, every site could be required to buy four gigabytes worth of magnetic disk. Users could ask the storage system to move a file they wanted to local storage from Bigfoot. Space would be automatically allocated on the local magnetic disk cache, and users would be notified (by email?) when their data was available. This space would be automatically garbage-collected. Users would be notified just before their data was removed from the cache.

There was also concern over the backup strategy for data stored on Bigfoot. We have no good ideas on how to back up ten terabytes of storage yet. Researchers with valuable data should keep copies locally (offline is fine) to protect against disasters.

7.2. Memory Problems

As mentioned in the visualization section, AVS uses excessive amounts of memory. Every site should load one or two Decstations or Alphas with 128MBytes of memory and use these for visualizing data under AVS. The project, spearheaded by Peter Kochevar, will work with the AVS developers to find a more elegant solution in the medium term.

7.3. Work on Alphas

Berkeley has an Alpha machine available for use by project members. Everyone on the project should have an account on it already. If this is not the case, contact jon forrest (jon_forrest@postgres.CS.Berkeley.EDU) to get an account. The machine is alpha.CS.Berkeley.EDU, and is currently running a beta version of OSF/1. For serious applications requiring a lot of computational time or memory, please coordinate with Jon Forrest to be sure that you have uncontested access to the machine.

Our agreement with DEC prohibits the publication of benchmark numbers for the machine. Nevertheless, researchers are encouraged to convert their programs to run on the Alpha, and to give us feedback on its performance. Jeff Dozier has agreed to set up and run a radiation model on the machine, and to compare its performance on a Decstation 5000.

7.4. Documentation

Peter Kochevar would like to see a glossary of Earth science terms for the computer science community. Hughes has published a list of acronyms and abbreviations for the EOSDIS project. Frew will try to get the electronic version of this list for augmentation by the project. Earth scientists with such glossaries are strongly encouraged to contribute them.

In addition, the project would like to collect manuscripts and technical reports from the Earth scientists. At present, all of the Sequoia technical reports have been produced by the computer scientists, sometimes with assistance from climatologists. We would like publications describing your work in the field.

8. Summary

The CRSEO trip was as successful as its predecessors. The major new contribution to the swat team's efforts was the definition of data lineage and computational steering. Project members at ESRG and Scripps are encouraged to send their views on these topics to Mike Olson (mike_olson@postgres.CS.Berkeley.EDU) electronically for inclusion in the benchmark specification.

Appendix E

Swat Team Trip Report

University of California at Los Angeles
Los Angeles, CA
Monday, 2 November 1992

1. Introduction

On Monday, 2 November 1992, the Sequoia 2000 Swat Team visited the University of California at Los Angeles. The meeting included swat team members (Jim Frew, Mike Olson, Peter Kochevar, Duane Waliser, and Allison Woodruff), Department of Water Resources meteorologist Frank Gerhke, and researchers and staff members at UCLA (Fabrice Cuq, Michael Ghil, Roberto Mechoso, David Neelin, Richard Turco, Bill Weibel, Ruiyan Lin, and Kayo Ide).

As was the case at CRSEO in Santa Barbara, researchers gave short presentations of their projects, and the group discussed visualization, data lineage, and computational steering.

2. Swat Team Trip Reports

Both Michael Ghil and Roberto Mechoso expressed concern over the detailed benchmarks described in previous swat team trip reports. For the very near term — before the January retreat — the benchmarks described are far too ambitious for implementation.

UCLA researchers would like to focus on two benchmarks before January.

First, it should be possible to combine and display two or three two-dimensional fields. For example, satellite images and field observations should be able to be displayed on a map. There should be different ways to do this visualization — superimposition of the images on the same canvas, displaying the datasets side-by-side on the screen, or laying one dataset on top of the other, with space between them, and letting the user see the bottom dataset by displaying the pair at an angle.

Second, it should be possible to manage and visualize four-dimensional data, and specifically spatial data over time. Slices, projections, and animation should be supported.

Olson stated that the swat team trip reports were intended to publicize the long-term goals of the project, and were not a commitment to implement all of the services they describe by January. A document describing the short- and longer-term goals of the project with respect to visualization, data lineage, and computational steering will be written after the swat team has visited all of the sites. This document will be available before 22 November 1992.

3. Visualization

In addition to the two short-term services described above, climatologists at UCLA would like other services from the data management and visualization software. This section gives a brief description of the major services desired.

It would be useful to have simple generic processing tools built in, such as bandpass filters and image operators. Image operators implement, for example, subtraction of images — pixels in one image are subtracted from pixels in another, producing a new image.

Intercomparison of datasets (such as displaying them on the same map) requires a tool to do projection conversion. All four of the sites visited have now requested such a tool.

UCLA researchers would prefer not to be required to use a single technique for doing space-variant data interpolation, such as optimal interpolation, since space-variant interpolation is an area of active research in the Earth science community. Any interpolation tool must be extensible.

Richard Turco urged the group to come up with an icon-based, Mac-like visualization tool that is easy for the scientists to use without consulting a computer programmer.

Any visualization tool should be able to produce images for use in conference presentations and in publications. The best interface would be for the scientist to be able to press a button to get a publishable figure from the screen image, and another to get a slide for a talk. The tool should permit citations and titles to be appended. Some default output format for these other media — black and white contour plots standing in for color maps, for example — should be provided.

Finally, researchers at UCLA were concerned with how they could get experience with the tools available under Sequoia. Peter Kochevar will return to Los Angeles for a one-day visualization session on 24 November, and may bring other members of his group from the San Diego Supercomputer Center.

At the November 24 meeting, Peter will use AVS to visualize some of Mechoso's GCM data. Both Peter and Frew got pamphlets describing the format of the GCM data. Peter's work with John Roads at Scripps is intended to give the project experience with visualization and an initial working system, which can be modified and extended for more general use.

4. Data Lineage and Computational Steering

UCLA researchers agree that the services described in the CRSEO trip report are interesting. They caution that for the short term, the project should concentrate on simple visualization benchmarks.

5. Data Management Issues

Duane Waliser has loaded a good deal of Coastal Zone Color Scanner (CZCS) data into Bigfoot under the Inversion file system. Very little tabular data of interest to people at UCLA has been loaded into POSTGRES to date.

Two important issues with respect to data management are standardization of formats and how and where to load data. Jim Frew chairs the sequoia-schema working group to deal with formats and metadata management, and researchers at all sites are encouraged to participate in this effort. Selection of format must be driven by Earth scientists, not computer scientists. Frew is also able to allocate space for loading of data.

At the 24 November meeting with Peter Kochevar and his group, UCLA researchers will discuss what they need in terms of data management and data loading.

5.1. Formats, Operators, and Data Loading

Roberto Mechoso asked for guidance on what should be done with data that is currently ready to load onto Bigfoot. For example, extant GCM output and observational data are in different projections and are at different resolutions. The Earth scientists are willing to convert data formats to whatever is in use in the project, but they do not want to have to do this conversion more than once.

Frew gave the following response:

- In the short term, researchers should treat the Bigfoot data store as a big file system. Data should be loaded in whatever format is currently available. The intent is to permit people to use the storage system to do work immediately.
- In the medium term, the computer scientists, in cooperation with the Earth scientists, will consider what operators and programs should be available to manage this data, and will add these to POSTGRES for use at all sites. Also, a standard format will be selected, and tools to do conversion automatically will be constructed.

Selecting a representation for scientific data is inherently an iterative process — formats will be selected and refined as their shortcomings are identified.

5.2. New Hardware

Mechoso and Michael Stonebraker were jointly awarded a grant by NASA to study high-performance computing systems. Their goal is to implement a parallel distributed global climate model on a network of Alpha workstations. This distributed computer would like to be a Bigfoot client, and any storage management protocol should consider it. The system will do modelling and visualization interactively, and will produce a timestep image every ten seconds of wallclock time.

Peter Kochevar and Jim Frew agreed to help Stonebraker and Mechoso draft a proposal for DEC requesting an equipment grant for Alphas to do the HPCC project.

6. Projects at UCLA

This section gives a very brief description of some of the major projects underway at UCLA.

6.1. Exploring Model Parameter Spaces

Both Michael Ghil and David Neelin are trying to explore the parameter space for complex models. The strategy is the following:

- Using some knowledge about physics, climatology, or a model's implementation, figure out what parameter variations are likely to be interesting. For example, if you change the solar constant or the length of the day, will the model do interesting things?
- Run a global climate model to verify that there's actually something interesting happening when you make the changes described. This requires computational steering and interactive visualization.
- After verifying that the GCM behaves as expected, try to analyze the behavior on a simpler climate model. Smaller models are easier to solve analytically, and so can give clearer understanding of basic physical processes than a complex global model.

Varying "constants" in this way is not as strange an idea as it sounds at first to computer scientists. For example, the rate of rotation of the Earth actually varies with time. The planet's rotation imparts kinetic energy to its atmosphere (the mountains push the air), so the planet slows down. Later, the air runs into other mountains, returning the energy to the earth. The variation in the length of the day can be on the order of milliseconds — a surprisingly long time, when you consider the inertial forces involved.

6.2. Visualizing Chemical Tracer Trajectories

Richard Turco and his students are working on the visualization of atmospheric flows. Model data are displayed over satellite images, and tracers are inserted in the image. The model output is then animated, and the paths of the tracers are displayed. At present, they're able to put a small number of tracers (fewer than ten) in an image. Eventually, they would like to do thousands.

The tracers are not simple points; they are volumes of air with internal state (for example, pressure and temperature). The visualization tool does not currently display this state, and they would like some guidance on how it might do so.

This is the first extensive visualization using AVS that the swat team has seen that was not done by Peter Kochevar. The person who did the AVS work at UCLA is Ruiyan Lin (ruiyan_lin@postgres.CS.Berkeley.EDU); climatologists interested in using AVS at other sites may be able to get some guidance from him.

6.3. Visualizing Atmospheric Flows

Christopher Strong and Michael Ghil are working on the visualization of atmospheric oscillation patterns over the Northern Hemisphere midlatitudes. The period of these oscillations is typically between 30 and 60 days. The data are obtained from a one-layer model and can therefore be animated by way of a 2-D sequence of images. The visualization package being used is Interactive Data Language (IDL); they like this package because of its ability to filter model data, perform power spectrum analysis, and compute statistics for a wide range of model or observational data. Finally, its animation module is easy to apply, and provides an advantage over traditional snapshots of model fields, allowing them to view the evolution of a selected pattern in real time.

In January at the Lake Arrowhead retreat, Strong and Ghil plan to present a three-way intercomparison of these 30- to 60-day oscillation patterns through a hierarchy of data extracted from

- the simple, one-layer model;
- the UCLA General Circulation Model; and
- real atmospheric observations.

Using visualization tools such as IDL, they plan to illustrate how animation is a powerful interactive tool in discerning the behavior of numerical models compared with that of the real atmosphere.

7. Miscellaneous

This section covers a side issues that arose during the day.

7.1. Neural Networks

NASA is funding another project at UCLA to recognize visual patterns in images. For example, one application would be to find Hurricane Andrew in a satellite picture of the eastern coast of North America. This would lend itself to content-based indexing of images.

Roberto Mechoso will see whether Sequoia can get the results of, or get involved in, this effort. A researcher at SDSC is looking for a good subject for a neural network application, and this may fit the bill. Kochevar will put this researcher in contact with Mechoso.

7.2. Goals of the Swat Team

An explicit goal of these swat team trip reports is to give researchers at all sites a detailed idea of who is active at other sites, and what projects they are working on. In particular, names and areas of expertise are included so that researchers can contact one another directly.

Project members are strongly encouraged to use these reports to get to know one another, and to develop ties with peers at other sites.

The final swat team report will describe the visualization benchmarks we will support in the short and long term. The trip reports will be included in that document as appendices.

7.3. Hardware Problems

UCLA purchased two input devices for their Decstation 5000s to be used with AVS. The devices consist of several knobs and a panel of pushbuttons. AVS on most platforms permits users to control visualizations using such an input device.

It turns out that AVS on the Decstations does not support this input device. Researchers at UCLA are disappointed. Frew will talk to Ira Machefsky and Peter Kochevar about this, so that both DEC and AVS are aware of the problem. DEC personnel with questions on the issue should contact Fabrice Cuq (fabrice_cuq@postgres.CS.Berkeley.EDU).

It is important that these devices be supported for visualization.

7.4. Alphas

Researchers at UCLA would like a FORTRAN compiler for the Alpha.

7.5. POSTGRES Client Software

Mike Olson will make a distribution package available to Sequoia sites that contains only the client software for POSTGRES. Sites need not copy and compile the full POSTGRES 4.0.1 distribution in order to use POSTGRES on heel.

8. Summary

The trip to UCLA resulted in a finer focus on short-term goals for visualization. Issues in data lineage and computational steering, while of interest to scientists there, were not considered as urgent as getting a good initial visualization tool.