How Sequoia 2000 Addresses Issues in Data and Information Systems for Global Change

Jeff Dozier
University of California
Santa Barbara, CA 93106-3060
dozier@crseo.ucsb.edu

Sequoia 2000 Technical Report 92/14
University of California
Berkeley, CA 94720

August 1992

Abstract

Sequoia 2000 is a project to design a next-generation information system for accessing, archiving, distributing, managing, and visualizing data for global change research. Funded by the Digital Equipment Corporation, it has investigators from computer science and Earth science departments on five campuses of the University of California.

1 Motivation and Objectives for Sequoia 2000

Global change researchers, particularly users of satellite remote sensing data, are seriously impeded by inadequate information systems. Users of remote sensing data need substantial knowledge about electromagnetic properties of the surface and atmosphere, and an ability to decipher arcane data formats. Data-intensive research in global change is also impeded by the immaturity of algorithms for estimating geophysical and biological variables from measurements of electromagnetic radiation from aircraft or spacecraft. Furthermore, agencies responsible for data focused too much on data acquisition, rather than storage, management, and distribution, hence the phrase "write-only archive" describes the status of many terabytes of historical satellite data[Dut90]. The study of global change has suffered because many of its brightest, most innovative practitioners have avoided the analysis of data and the use of data to verify and calibrate their models.

Hence the major objective of Sequoia 2000 is to develop an improved data and *information* system for global change researchers that fosters synergistic interactions between observations and models and enables and encourages interdisciplinary research.

2 Sequoia 2000 Infrastructure

Digital Equipment Corporation's flagship external research project, Sequoia 2000 has 22 investigators from computer science and Earth science departments on five campuses of the University of California[Ston92]. The Project Directors are Michael Stonebraker from the Berkeley campus on the computer science side and Jeff Dozier from the Santa Barbara campus on the global change side.

The following sections describe some of the global change and computer science areas of expertise. All investigators can be reached through the electronic mail address last_name@postgres.berkeley.edu.

2.1 Sequoia Global Change Researchers

Among the Sequoia co-investigators are numerous scientists with a history of significant discoveries of the behavior of the Earth System.

Atmospheric chemistry

Richard Turco of UCLA was one of the discoverers of the "nuclear winter" phenomenon in the 1980s and also contributed to our understanding of the ozone-destruction cycle in the stratosphere. His Sequoia investigation seeks to couple models of stratospheric chemistry with atmospheric circulation.

Cloud and Earth radiation budget

Catherine Gautier of UC Santa Barbara examines the surface radiation budget as it is affected by clouds. She uses data from the NOAA polar orbiting and geostationary satellites, and is a Team Member on the Advanced Infrared Sounder (AIRS) on the EOS satellites.

Global climate models

Carlos R. Mechoso of UCLA is one of the leading developers and analysts of global climate models. His current research concerns coupling the circulations of the atmosphere and oceans.

Cloud-climate interaction

Richard C. J. Somerville of Scripps Institution of Oceanography (UC San Diego) examines the radiative forcing of climate (e.g. surface convection) by clouds.

Hydrology

Jeff Dozier (UC Santa Barbara) examines the links between hydrology and hydrochemistry of snow-covered alpine basins in the Sierra Nevada, Rocky Mountains, Austrian Alps, Tien Shan of China and Kazakstan, and Andes. He is an interdisciplinary principal investigator on EOS, and returns to UCSB this fall after a two-year appointment as the EOS Senior Project Scientist.

Ocean ecology and productivity

Raymond C. Smith (UC Santa Barbara) studies ocean productivity using satellite data. He has spent the last two winters in the Antarctic Ocean examining the effects of increased ultraviolet radiation, caused by the ozone hole, on ocean phytoplankton.

Regional climate models

John Roads (UC San Diego) studies the coupling between global climate models and climate over large regions, e.g. the western U.S. He particularly examines the effects of climate change on precipitation.

Terrestrial ecology

Frank W. Davis (UC Santa Barbara) uses satellite data to study vegetation patterns and their relation to topography and climate. He has also used tree ring data to interpret past climate conditions in California.

2.2 Sequoia Computer Science Researchers

Among the Sequoia co-investigators are numerous computer scientists who have contributed to significant innovations in software and hardware.

Tertiary storage

Randy Katz and David Patterson (UC Berkeley) developed RAID (redundant arrays of inexpensive disks) technology, and are applying these same ideas to other kinds of storage media.

Tertiary file systems

Carl Staelin and Tom Anderson (UC Berkeley) are developing new file systems to address tertiary storage, and will compare these to commercially available systems, line UniTree and Ninja.

Networking

Joseph Pasquale and George Polyzos (UC San Diego) are developing new network protocols for fast transfer of large data sets. Included are modifications to the

Unix operating system, as well as special protocols for guaranteed delivery and for transmission of real-time video data.

Data base management

Michael Stonebraker (UC Berkeley), developer of the commercially available Ingres data base management system, has invented a next-generation DBMS, Postgres, that incorporates spatial queries and the concept of large objects.

Geographic information systems

Kenn Gardels (UC Berkeley) has coupled the GRASS geographic information system to Postgres. Under development is the extension of Postgres to handle GIS types of queries.

Scientific visualization

Peter Kochevar (DEC) and Michael Bailey (San Diego Supercomputer Center) leads the visualization effort to couple existing visualization tools (IDL, AVS, and Khoros) to Postgres, and to address scientific requirements for a new visualization system.

Electronic repository

Ray Larson (UC Berkeley) addresses the inclusion of textual data in Postgres, including full-text retrieval and queries.

3 Overview of Global Change Research

We know from the geologic record that the Earth has undergone significant alterations in its physical configuration since the planet was formed some 4.5 billion years ago. Indeed, the entire history of the Earth is a dynamic continuum, with changes on all time scales. This evolving character of the Earth makes its scientific study both interesting and difficult. What is disturbing, however, is evidence that the rates of change for several key components of the system are increasing dramatically, and that human activity is responsible. Within the span of a single human lifetime, we can detect significant differences in ecology, atmospheric chemistry, terrestrial and oceanic ecosystems, and water quality. Whether human society affects the Earth system is no longer a question. The question has now become, "How severe is that effect and what can be done about it?" The table below summarizes some aspects of the global change problem.

Table 1: Some Causes of Global Change

Problems	Specific Facts
Increasing greenhouse gases	More than 5 Gton/year of carbon into atmosphere from fossil fuels
Decreasing stratospheric ozone	5 %/year increase in chlorofluorocarbons
Decreasing biodiversity	Decreasing sizes of habitats
Deforestation	More than 10 million hectares/year
Burgeoning human population	Projected more than 6 billion by 2000

3.1 Accelerated Change in Earth System Caused by Human Activity

Continued human-induced change in the Earth System appears inevitable. Increases in atmospheric concentrations of carbon dioxide can be tied directly to combustion of fossil fuels, deforestation, and other activities that are not likely to decrease soon[IGBP86]. The following factors also contribute to a picture of continuing global change driven by human activities:

- Increases in world population, the strong association between standard of living and energy consumption, and economic factors linked with the abundance of fossil fuels suggest that carbon emissions into the atmosphere will increase for the foreseeable future.

- Even if the countries of the world agreed today to freeze emissions of carbon dioxide at current amounts, the atmospheric concentration would continue to increase. Stabilizing atmospheric CO_2 at the current level, which is 25% above the pre-industrial concentration, would require an unrealistic 60–80% reduction of carbon emissions.

- Population increases define land use changes and agricultural production in ways that contribute to global change. For example, rice cultivation and cattle production are significant sources of methane, and other important radiatively active gas in the atmosphere.

- Chlorofluorocarbons manufactured for refrigeration and air conditioning break down when they reach the stratosphere, and produce the chlorine radical that destroys ozone. The chlorine radical has a 100-year residence time in the upper atmosphere, so we are only beginning to see the significance of human-induced changes in the chemistry of the stratosphere.

Unfortunately, future global changes will not be limited to the causal factors we have already identified. Subtle, yet significant, changes that are now undetected will also occur. Thus, the inevitability of global change and the prospect of unanticipated modifications of the global environment add urgency to our need to understand the operation of the Earth System.

3.2 Implications

Societal implications of human-induced global changes are enormous. Global climate models agree that a doubling of carbon dioxide concentrations in the atmosphere, as compared to pre-industrial concentrations, will result in a 2 – 4°C globally-averaged warming of temperatures in the troposphere. The last global temperature change that large—a decrease instead of an increase—occurred during the Ice Age advance and retreat, and brought glacial ice within the boundaries of what is now the United States. However, the Ice Age advance/retreat took place over tens of thousands of years, whereas these human-induced changes are projected to occur within a century. Atmospheric CO_2 will rise to concentrations substantially greater than any in the last 160,000 years, as evidenced by measurements in air bubbles trapped in ice cores from Antarctica and Greenland.

Global warming is not the only result of changing atmospheric composition. The altered circulation associated with shifts in atmospheric temperature will also affect water balances. Some climate simulations show an increased frequency of drought in many mid-latitude continental regions. Such droughts would reduce water availability and quality for industry, agriculture, and human consumption. Warming may also melt portions of the polar ice sheets and cause the sea level to rise, threatening coastal cities and environments.

Limitations of time and practicality make it difficult to ameliorate global change soon. However, if we can improve our predictions of the plausible consequences of future changes, that will provide society with the opportunity to adapt and minimize negative effects.

3.3 Challenge for Earth Systems Science

In addition to integrated measurements of Earth system processes, we must also have an integrated approach to our understanding of the resulting information. In general, the diverse disciplines that comprise the Earth sciences developed independently, and they have tended to pursue independent research objectives and strategies. Developments in observational methods, theories, and models in the fields of meteorology, oceanography, bioclimatology, ecology, geochemistry, geomorphology, and hydrology progressed primarily because of advances within each discipline. Now, however, three forces have combined to alter the modes and focus of research about the Earth:

1. The maturity of traditional disciplines has brought the research frontiers to the interfaces of the Earth Science disciplines.

2. The new view of the Earth from space encourages an integrated approach that sees the planet as a whole system.

3. The growing awareness and apprehension about the effects of human-induced global change makes interdisciplinary methods critical as well as desirable.

3.4 Elements of Global Change Research Program

Objectives of research in global change are to study, predict, and understand changes, both natural and anthropogenic, in the Earth System[CEES91]. This requires improvements in observations, models, and information systems. Sequoia 2000 addresses information systems, with requirements driven by global change researchers at four University of California campuses. Within Sequoia 2000, the global change researchers address the following key problems identified by the Intergovernmental Panel on Climate Change[IPCC90]:

- the "greenhouse effect," associated with increasing concentrations of carbon dioxide, methane, chlorofluorocarbons, and other radiatively active gases;

- ozone depletion in the stratosphere, resulting in a significant increase in the ultraviolet radiation reaching the Earth's surface;

- changing water chemistry in alpine ecosystems caused by changing precipitation chemistry;

- deforestation and other anthropogenic changes to the Earth's surface, potentially affecting the carbon budget, patterns of evaporation, precipitation, soil erosion, and other components of the system;

- changes in photosynthesis, respiration, transpiration, and trace gas exchange, both on the land and in the ocean;

- decline in the health of vegetation, because of long-term changes in the chemistry of the atmosphere, precipitation, runoff, and ground water.

Improved information systems will allow Earth System scientists to study processes and develop better models. Better, more comprehensive information systems are critical to the success of the study of global change over the next two decades, so that researchers can work with scientific information instead of satellite radiances.

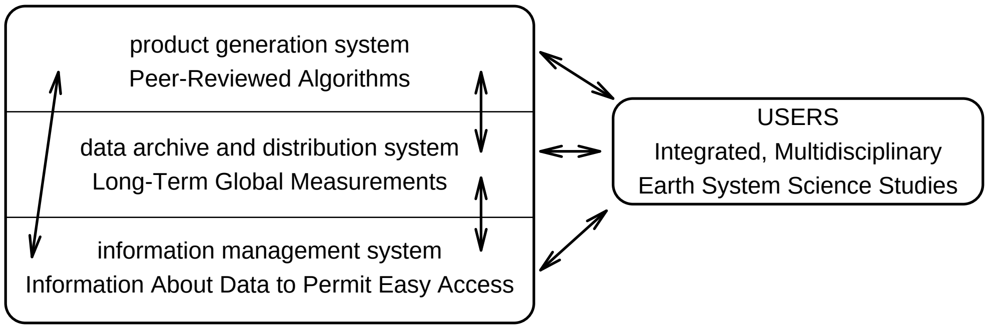

Figure 1: Conceptual View of an Information System for Global Change

4 Earth Observing System (EOS)

NASA's part of Global Change Research Program is the Earth Observing System, a long-term project extending through 2013[Doz90]. Precursor missions began with the launch of the Upper Atmosphere Research Satellite (UARS) in 1991. The core of the EOS program consists of 6 satellites with about 20 sensors launched between 1998–2002, with instrument teams to develop scientific products[Moor91]. The volume of raw data and scientific products will exceed 1 terabyte per day by 2000. There are 29 interdisciplinary investigations—one Sequoia co-PI is also an interdisciplinary PI—and 570 named co-investigators, including one Sequoia co-I[NASA91]. The total user population of EOSDIS is estimated at 10,000.

4.1 Relation of EOS and Sequoia 2000

Sequoia 2000 is a research project linking computer science and global change research[Ston92]. The EOS Data and Information System (EOSDIS) includes operational system and next-generation prototypes, but because of their requirement

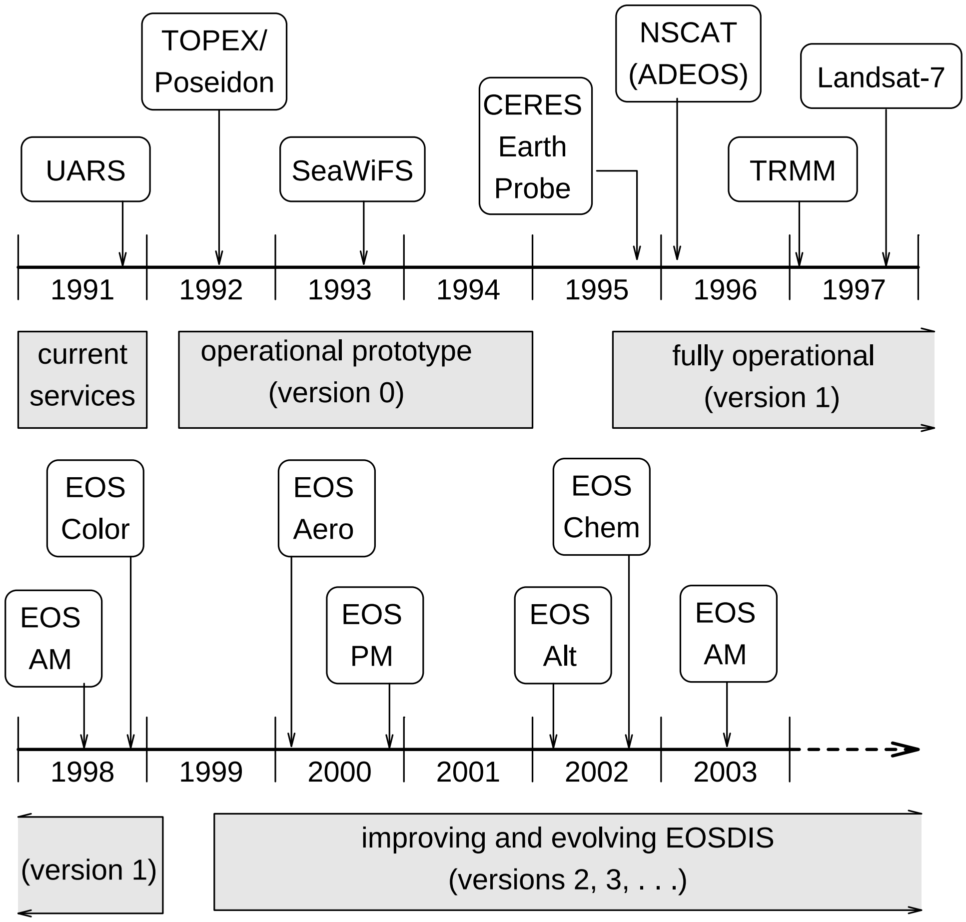

Figure 2: Schedule of U.S. Satellite Launches and EOSDIS Versions

to deliver soon an operational system, the prototypes are conservative. EOSDIS Version 0, providing access to existing operational satellite data, is scheduled for availability to users in 1994[Doz91]. Features to be incorporated include:

- interoperability between EOS archive centers;
- science processing library;
- ingestion of Pathfinder data sets (scientific products developed from current operational sensors).

Prototypes developed in Sequoia 2000, described in Section 5, will be available for incorporation into EOSDIS.

5 Issues in Information Systems for Global Change

5.1 Data Base Management

Current data base management systems are too primitive for global change research, hence they have not been adopted by the science community to the extent, for example, that modern numerical methods are incorporated in scientific models. Existing data base management software is typically designed for industrial applications with a huge number of short records, but where the records are characters or numbers. Queries on mature systems are usually optimized.

However, the paucity of data types and the absence of geographic queries reduce the usefulness of common data base management systems. Earth System scientists want to query and analyze:

- raster data (e.g. satellite images, digital elevation grids)
- vectors
    - polygons (e.g. drainage basin boundaries)
    - directed graphs (e.g. river networks)
    - simple vectors (e.g. profiles of atmospheric humidity)
- point data (e.g. river discharge records)
- text data (e.g. algorithm descriptions, processing histories)

The challenge faced by Sequoia 2000 is to develop or modify a data base management system that satisfies needs for remote sensing, geographic information systems, diverse data types, and operations. These require different query indexing and access methods.

5.2 Physical Data Organization

Scientific problems faced by global change scientists require fast network access to huge amounts distributed, heterogeneous data. The volumes are large— stored satellite data, derived products, and model output will add up to 500 terabytes/year by 2000—and the objects themselves are large. For example a single Landsat Thematic Mapper image is 300 megabytes.

The challenge faced by Sequoia 2000 is to develop:

- a file system and database management system that understand distributed, tertiary memory;
- a data base management system that efficiently indexes and accesses large objects (including clever abstracts);
- a system that optimizes performance both for users' models and users' queries.

5.3 Sample Queries

Some sample queries illustrate the problems we face:

Raster queries

- "Select all day-time AVHRR images for the Southern Sierra Nevada between October 1991 and June 1992 and sort chronologically."
- "Select
  - band 1 $\geq$ shadow threshold (anything less is not snow)
  - band 5 $\leq$ cloud threshold (gets rid of clouds)
  - $\frac{\text{band } 4 - \text{band } 5}{\text{band } 4 + \text{band } 5}$ [snow grain-size index]

  for Landsat Thematic Mapper image covering Bubbs Creek area nearest to April 1, 1992."

Directed-graph queries

- "Identify all reaches of the San Joaquin River above Friant Dam."

Polygon queries

- "Show land use and land cover within 500 km of Ürümqi."

Point queries

- “Show all snow courses in the Kern River basin and tabulate their snow accumulation in April 1992.”

Joins between types

- “Show ocean phytoplankton concentration underneath the ozone hole in an ice-free area for dates nearest October 15, 1991.”
- “Show all snow courses in the Kern River basin having more than 500 mm of snow water equivalence in April 1992.”

5.4 Development of Applications

Science software must interface with the data base management system. Model codes are created by Sequoia scientists. The require:

- interface between algorithms and data base management system;
- libraries for common functions;
- easy import of user-supplied algorithms.

The challenge is to provide an system whereby ideas are more easily translated into code that can access the data. Access to data should be provided through the data base management system, instead of through files. The DBMS should be designed for easy “recipe management” (e.g. boxes and arrows paradigm).

5.5 Remote Visualization

we remotely visualize large data sets stored on a multi-terabyte storage system. We render data on local workstations, so that we can: browse through many compressed images; manage, use, and manipulate large data sets through a data base management system; and use “computational steering” to guide models.

Challenges for the visualization system are:

- it must handle large objects (single images of several hundred megabytes);
- it must exploit the capabilities of a distributed, heterogeneous computing environment, including workstations, vector and parallel supercomputers, and networks;

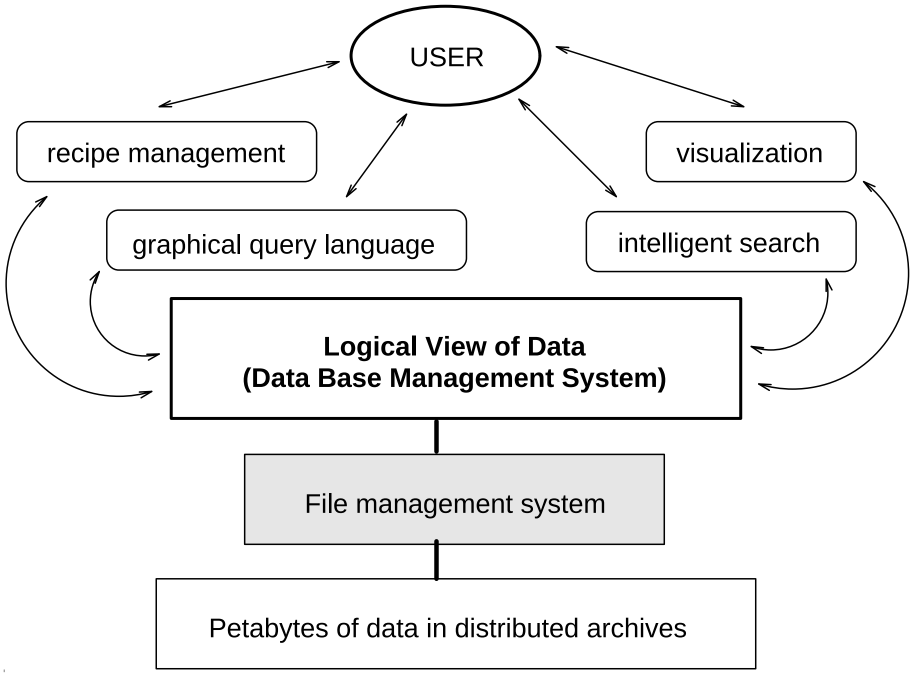

Figure 3: "DBMS-centric" View of Scientists' Interaction with Data

- it must provide interactive I/O with models, so that changes in model output, caused by interactively changing parameters, are immediately displayed;
- the data base must be queried using graphs, maps, and images (as well as text).

To support these functions, we couple the visualization packages (we support IDL, AVS, and Khoros) to the data base management system (Postgres). We provide interfaces to a wide variety of data sets, and output to a spectrum of hard-copy devices. Also coupled to the data base system are a graphical query language and a recipe manager.

5.6 Networks

Current applications in Earth System science bring networks to their knees. The success of a system to provide access to many terabytes of data in distributed archives hinges on networks. Sequoia 2000 researchers will be connected by a 45 megabit/sec network.

The challenges faced by the network component are to develop new protocols that address needs for particular applications, e.g. guaranteed delivery for reserved periods. Moreover, the operating system must not be the bottleneck; fast networks are expensive, and we want to make sure we use the band-width.

5.7 Data Base Administration

Scientists need to trust the functions for

- integrity (e.g. time variation of calibration coefficients);
- concurrency and triggers (processing of scientific products depends on availability of input data);
- version control and tracing (algorithms change).
- distribution, replication, and backup.

The challenge is that the system must provide “data lineage,” a reproducible audit trail for every item. It must also provide information about algorithms and functions, as well as data objects, and must prevent invalid operations on data.

5.8 Evolution

As Sequoia 2000 develops, it must not build in obselesence. Requirements are driven not only by changing computing technology (processing, storage, networks, software), but also by changing science (knowledge and algorithms) and changing scientists’ behavior and familiarity with computing.

Sequoia 2000 (and EOSDIS to a much larger clientele) provides operational services, yet must allow for evolution. It must incorporate advances in computing technology and not lock into obselesence. Information management must allow for insertion of next-generation DBMS. Implementation of parallel processing hardware might be constrained by difficulty of porting existing software. Both near-term and long-term goals must be accommodated.

Acknowledgements

This research was sponsored by the University of California and the Digital Equipment Corporation under Digital’s flagship research project Sequoia 2000. Other government and industrial partners include the California Department of Water

Resources, the U.S. Geological Survey, ESL, Hewlett-Packard, Hughes Aircraft Company, Metrum Information Storage, Research Systems Incorporated, and Science Applications International Corporation. My work on EOS is supported by the National Aeronautics and Space Administration.

References

[CEES91] Committee on Earth and Environmental Sciences. *Our Changing Planet: The FY 1992 U.S. Global Change Research Program.* Office of Science and Technology Policy, Washington, D.C., 1991.

[Doz90] J. Dozier. Looking ahead to EOS: the Earth Observing System. *Computers in Physics*, 4(3):248–259, 1990.

[Doz91] J. Dozier and H. K. Ramapriyan. Planning for the EOS Data and Information System (EOSDIS). In R. W. Corell and P. A. Anderson, editors, *Global Environmental Change*, NATO ASI Series I: Global Environmental Change, Vol. 1, pp. 155–180. Springer-Verlag, Berlin, 1991.

[Dut90] J. A. Dutton. The EOS Data and Information System: concepts for design. *IEEE Transactions on Geoscience and Remote Sensing*, 27(2):109–116, 1990.

[IGBP86] U.S. Committee for an International Geosphere-Biosphere Program. *Global Change in the Geosphere-Biosphere: Initial Priorities for an IGBP*. National Academy Press, Washington, D.C., 1986.

[IPCC90] J. T. Houghton, G. J. Jenkins, and J. E. Ephraums, editors. *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, Cambridge, 1990.

[Moor91] B. Moore, J. Dozier, M. R. Abbott, D. M. Butler, D. Schimel, and M. R. Schoeberl. The restructured Earth Observing System: instrument recommendations. *Eos, Transactions, American Geophysical Union*, 72(46):505, 510, 516, 1991.

[NASA91] NASA. *EOS Reference Handbook*. NASA Goddard Space Flight Center, Greenbelt, MD, 1991.

[Ston92] M. Stonebraker. An overview of the Sequoia 2000 project. In *Proceedings 1992 COMPCON Conference*, San Francisco, 1992.