The First Cedar Data Base Report

R.J. Sica, Center for Atmospheric and Space Sciences, Utah State University, Logan, UT
A.D. Richmond
B.A. Emery
S. Chakrabarti, Space Sciences Laboratory, University of California, Berkeley, CA
V.B. Wickwar, SRI International, Menlo Park, CA
The Coupling, Energetics and Dynamics of Atmospheric Regions (CEDAR) program is an international, cooperative effort within the aeronomy community to further our understanding of Earth's upper atmosphere. This program will be highlighted by numerous multi-instrument measurement campaigns and global simulations of the magnetosphere-ionosphere-atmosphere system. To achieve the full scientific benefits of these observations and simulations will require the ability to access large amounts of data from many different instruments and models. This report describes a recommended data base system for the CEDAR program to help establish this access.

Several general principles help to determine the most desirable form for the CEDAR Data Base. Data from the CEDAR program will be collected primarily for research purposes, and much of it will be state-of-the-art, that is, subject to evolving experimental uncertainties and problems, requiring active involvement of the experimenters in its interpretation. The data base must therefore be more than a collection of data, and must take advantage of whatever relevant skills are available in the community. It must be able to handle rapidly expanding data rates, and provide quick and easy access to well-documented data. Specially developed software for data management and analysis should be shared among institutions and scientists. Commercially available software should be taken advantage of when it offers significant advantages in the display and management of data. High-density forms of data storage are becoming available that can offer advantages over 9-track tapes for transferring data. A central data base facility is desirable to act as a focus of data base efforts, to store and make available widely used data and software, and to provide user services. Rather than raw data, the central facility should store basic and derived geophysical parameters, in order to make the data easy to use. Raw data should be made available directly from the experimenters. Data access through national and international computer networks, as well as by copies on magnetic tape and other media, will be needed. The CEDAR Data Base should also link to relevant information in related data bases maintained by NASA and other organizations. It should help facilitate basic information exchange in the forms of electronic mail, up-to-date lists of scientists, and catalogues of data.

The responsibilities for the CEDAR Data Base should be shared among four principal groups, whose members often overlap: the CEDAR Data Base Committee, the central facility, the data suppliers, and the data users. The Data Base Committee should be the principal coordinating body among the aeronomy community for determining types of data to be included, establishing common data formats and schedules of submission,
and for recommending user interfaces with the data base, in terms of access modes and services to be provided. Subcommittee reports addressing these issues are included here as appendices. The central facility should maintain desired data collections and related information, provide catalogue and other information to interested users, collect and develop useful software for data management and analysis, make data and other information accessible for scientific research through mailed copies and online facilities, promote contacts between data users and data providers, and host scientific research using the data. The data suppliers need to help the Data Base Committee determine appropriate data and formats; to process, verify, and document the data and submit it to the central facility; and to interact with users of the data. The data users need to interact with the data suppliers to assure appropriate uses of the data, and to offer coauthorship on publications, and they need to keep the central facility informed of uses made of the data.

Several recommendations are made concerning the implementation of the CEDAR Data Base. The Data Base should be built upon the existing Incoherent Scatter Radar Data Base with the central facility at the National Center for Atmospheric Research (NCAR). The data format currently used for the incoherent scatter data should be adopted for most other types of CEDAR data, although the NASA Standard Format Data Unit (SDFU) should be considered in the future for some applications. The central facility should attempt to obtain a dedicated minicomputer to carry out the online data base functions. This computer should use the UNIX operating system. A portable version of the Millstone Hill data base system would be desirable as a basis of the online system to be developed at the central facility. The central facility should have good connections to the SPAN and NSFnet networks. The Rules of the Road currently in effect for the Incoherent Scatter Radar Data Base should be considered for adoption for the CEDAR Data Base, allowing users broad access to data at the central facility, but requiring them to inform data suppliers about intended uses and to offer coauthorship on publications. The National Science Foundation should support the CEDAR Data Base as a central element of the CEDAR program, and as a long-term community effort. Looking to the future, the CEDAR Data Base should keep on top of relevant developments in distributed data base techniques, telescience techniques, and artificial intelligence.
# Table of Contents

**Executive Summary** iii

**Table of Contents** v

**Introduction** 1

Why does the CEDAR program need a data base? 2

What kind of data base does the CEDAR program need? 4

- *Basics of data bases* 4
- *CEDAR Data Base specifics* 5

**Organization of the CEDAR Data Base** 10

- *Responsibilities of the CEDAR Data Base Committee* 10
- *Responsibilities of the central facility* 12
- *Responsibilities of the data supplier* 13
- *Responsibilities of the user* 13

**Recommendations for implementing the CEDAR Data Base** 14

- *Advantages of merging with the Incoherent-Scatter Radar Data Base* 14
- *Data Formats* 14
- *Data base computer* 16
- *"Rules of the Road" and security concerns* 17
- *The bottom line* 18

**Long range objectives of the CEDAR Data Base** 19

**Acknowledgment** 19

**References** 20

**Table 1** 21
THE FIRST CEDAR DATA BASE REPORT

Figures

Appendix A: The CEDAR Data Base Committee

Appendix B: Reports from the Group Leaders

Incoherent-Scatter Radars by V. B. Wickwar
Spectroscopy Data Base Requirements by C. S. Deehr and S. Chakrabarti
Middle Atmosphere Radars by A. Manson
Contributions of Optical Interferometers to the CEDAR Data Base by C. A. Tepley
Considerations for an Imaging Component of the CEDAR/NCAR Data Base by M. Mendillo and J. Baumgardner
CEDAR Lidar Data Base Requirements by C. S. Gardner
Incorporation of Modeling Results in the CEDAR Data Base by C. Fesen

Appendix C: An example of the NCAR tape format

Appendix D: Proposed CEDAR Data Base “Rules of the Road”

Appendix E: Astronomical data analysis from remote sites

by S. Chakrabarti, C. Dobson, G. Kaplan, H. Marshall, M. Lampton
R. Malina, S. Bowyer and J. Star
In the last five years the United States aeronomy community has organized a cooperative program to investigate the physics and chemistry of the Earth’s upper atmosphere. The Coupling, Energetics and Dynamics of Atmospheric Regions (CEDAR) program has promoted multi-instrument and multi-model campaigns, which include international participation, to attack aeronomic problems too complicated for single ground-based stations or lone satellites to address. A summary of this program has been presented by Romick et al. [1987]. Details of the implementation of the CEDAR program is contained in Volume 1 of the CEDAR report to the National Science Foundation [1986]. The community realizes that an essential part of the CEDAR program that will involve all its participants is the CEDAR Data Base. To facilitate the exchange of large amounts of experiment and modelling data that will be generated by various CEDAR studies, a well-organized structure to handle these data is necessary. This report is an overview of the CEDAR Data Base Committee’s initial finding on the need for a data base, what kind of information should be included in the data base, preliminary descriptions of the responsibilities of all parties, and some short and long range plans for the data base.

The CEDAR Data Base Committee has two major functions. The first function is to design guidelines for submission of data to the data base (i.e. input). Input Group Leaders (Appendix A) have been chosen to represent each major discipline of the CEDAR program. The Input Group Leaders are responsible for discussing with individual scientists in their discipline the data base needs of each area, i.e. what measurements from an instrument or output from a model should be in the data base, what format should be adopted for the data, how often should data be contributed, etc. At the time this document is being written many of these groups are just beginning discussions of these and other important topics. Initial reports from the Group Leaders are in Appendix B.

The second major function of the Data Base Committee is to provide advice for the development and access of a data base system for the contributed measurements (i.e. output). Activities of the Committee in this area include investigating what types of
software are necessary for a successful data base and to ensure easy accessibility for users. In addition to the Input Group Leaders, the Data Base Committee includes members interested in issues related to the access and use of the data (Appendix A).

2. WHY DOES THE CEDAR PROGRAM NEED A DATA BASE?

The term data base as used in the context of the CEDAR program is broader than the common usage describing a class of computer codes which handle data structures. Data base activities include developing the infrastructure required to obtain measurements and model outputs from a variety of users in a predetermined format, to manipulate the data they have into a format that can be easily accessed by the CEDAR community, and to make this data and critical ancillary information available to CEDAR participants. Cooperative interactions among many institutions and scientists will be required for the data base to operate effectively. Not only data but also software for data management and analysis needs to be shared. It is envisioned that a central facility will act as a focus of data base activities, but that responsibilities will be shared among the CEDAR Data Base Committee, the central facility, data suppliers, and data users.

The amount of information on the Earth's upper atmosphere has grown exponentially since the International Geophysical Year (IGY) in 1958-1959. The projected growth rate of NASA Solar and Space Physics data as presented in the CODMAC [1986] report on NASA data base issues is shown in Figure 1. Millions of federal dollars have been spent on the acquisition and analysis of aeronomic data. The measurement campaigns being planned for the CEDAR program promise another explosion in our knowledge of the upper atmosphere similar to the IGY. As scientists we have an obligation to carefully document and archive our work for all researchers to use. Therefore, records of our experiments and models must be carefully archived. This archive is one of the important purposes of a community data base. Only wise management practices instituted now will be able to control the information explosion which faces the CEDAR program.

A few examples of the large amounts and complexity of data to be managed will clarify these points. The CODMAC [1986] report gives a data rate of $3 \times 10^{11}$ bits/year for the
Dynamics Explorer satellites. There are presently 6 incoherent-scatter radars in operation, which routinely measure in space and time several ionospheric parameters: electron densities, ion velocities, electron and ion temperatures, ion composition, and ion-neutral collision frequency. A typical incoherent-scatter radar, such as that at Sondrestrom, Greenland, produces about $10^{11}$ bits/year (Wickwar [1984]). Fabry-Perot interferometer measurements of the oxygen red line, which yield neutral wind and temperature, produce about $10^9$ bits/year. Presently at least 20 of these devices are in operation, contributing another $2 \times 10^{10}$ bits/year. Other basic aeronomical devices include photometers, imaging systems, lidars, and spectrometers/spectrographs that will also generate large amounts of data in many different forms, coming from many different institutions and involving many different researchers. Managing the total CEDAR Data Base is a task equivalent to managing the data base system for several satellite programs. Furthermore, if large amounts of output from models of the atmosphere and ionosphere are included, in addition to ancillary geophysical parameters, a vast amount of information needs to be ordered.

Past measurement campaigns, together with future CEDAR programs, have obtained and will obtain measurements in unique geophysical situations with probability of reoccurrence on the order of tens of years or more. An example of this is the Great Red Aurora of February 11, 1958. Due to the foresight and diligence of the IGY scientists, priceless auroral measurements were obtained that are still being utilized today. Carefully documented and archived data is a valuable resource for future scientists, since the value of a particular set of measurements is often not fully realized at the time they are acquired. Careful documentation is important for scientists analyzing present data, so they can be aware of various peculiarities in the data.

With the evolution of research in upper atmospheric science, the solution of new problems increasingly requires access to multiple types of data, or to a particular type of data obtained at multiple sites. The number of problems that can be successfully attacked with single data sets is diminishing. In order to promote scientific progress, data from many different types of instruments and from many different sites must be made available to researchers. An organized data base structure is necessary to facilitate the easy access to
and analysis of these data.

3. WHAT KIND OF DATA BASE DOES THE CEDAR PROGRAM NEED?

3.1. Basics of Data Bases

There exist four basic types of data which can be included in a scientific data base. These types are raw data, basic data, basic (geophysical) parameters and derived parameters. An example of these data types for incoherent-scatter radar data is shown in Table 1, from Wickwar [1984]. Raw data is the basic information that an instrument produces, and is usually recorded on magnetic disk or tape. Raw data, to be useful, must include instrumental parameters and calibrations necessary to reduce a given set of measurements. The raw data is typically processed, by co-addition or re-scaling, to yield the basic data from the instrument. Parameters of geophysical interest derived directly from the basic data are the basic parameters of the instrument. Derived parameters are the next level of data processing. The calculation of derived parameters may include the use of a model that must add additional assumptions to the analysis. An optical example of a derived parameter is the determination of the incoming electron energy spectrum from the ratio of optical emissions. Outputs of global computer simulations would be another example of derived parameters. The term derived parameters will also be considered appropriate to classify output from “first-principle” models. It is, therefore, an issue unto itself to decide which, if not all, levels of data from a given instrument should be part of the data base.

The data base itself can be centralized or distributed. A centralized data base relies on community access to a single machine (or cluster of machines) at one physical location. An example of this is the NCAR Incoherent-Scatter Radar Data Base. A distributed data base can be of one of two types. In the most extreme case of a distributed data base, each individual user has all of the data at his or her own site. The Lenhart tape of geomagnetic and solar parameters, distributed by the World Data Center, is such a data base. Another type of distributed data base is one where different types of data reside at different locations and one central facility contains summary data from all of the instruments. The CEDAR
Data Base Committee recommends that distributed data base systems be integrated with a strong centralized facility.

Opinions on the above topics and other data base concerns were solicited from the aeronomy community in 1986 in a written survey distributed by A. Richmond, head of the Incoherent-Scatter Radar Data Base. Responses were received from 35 scientists, 12 from foreign countries. The responses stressed the importance of data documentation, clean data and the existence of easily browsed summary data. Though derived parameters were perceived as important, most respondents felt that it would be most appropriate to generate these on the data base system as needed using existing software. Users also stressed that the ability to search and select data using user-specified criteria on the data base computer was important. There was a general consensus that a centralized data base was highly desirable, although many foreign scientists were rightly concerned about access to a remote, centralized data base. A major area of agreement was the importance of having knowledgeable people available to help understand the data and access the system.

3.2. CEDAR Data Base specifics

There are many issues to be considered on the formation of a data base that best fulfills the needs of the CEDAR community. Many of the following topics have been discussed at previous CEDAR meetings and among members of the Data Base Committee. However many of these topics are still open questions that need community input.

In order to be most effective in utilizing the capabilities available in the CEDAR scientific community, and to be most cost-effective in obtaining the needed CEDAR Data Base capabilities, the data base functions should be shared among the institutions involved. Software developed at one facility should be made available to other facilities for local use or access via a network; portability should be a high priority when new software is developed. The data base system developed at the Millstone Hill Observatory (Foster et al. [1985]) would be quite valuable if it were made portable. Many other institutions have also developed useful programs for data analysis that should be shared. Some large data sets can be maintained most cost-effectively at individual sites, and made available on request.
There is, however, an important need for a central facility to act as a focus for data base efforts, to provide a location where at least summary data from all instruments can be obtained, to organize and make available catalogues of the data, to provide a software library of useful analysis programs, and to provide computing facilities for users who have only limited facilities available. The presence of a central facility will strongly encourage conscientious participation in data base activities by those collecting the data: in order to be able to submit data and have it catalogued the data providers will have to address and solve all the problems of data interpretation, reduction, reformatting, and documentation. In order to interact effectively with the providers and users of data, the central facility should have scientists actively involved in research with the data. As described later in this report, the NCAR Incoherent-Scatter Radar Data Base has the appropriate characteristics to form the basis of the CEDAR Data Base Central Facility.

In the previous section the four types of data were defined and discussed. Should the CEDAR data base contain all four types of data or only some types? Since the vast majority of users are only interested in geophysical parameters (basic and derived), the data base should contain at least these data types. If experimental derived parameters are not in the data base, the necessary programs to compute these quantities should be available to the user. Opinions vary on how to obtain derived parameters from the basic parameters and various assumptions. Hence, there should also be adequate documentation on how each site derived these parameters as well as “generic” programs available at the data base to regenerate them. However, there are occasions when it is extremely useful to look at the raw or basic data, especially before standard analysis procedures are implemented for each instrument. The raw data allows a user to re-analyze measurements if they desire. The basic data is necessary to allow the user to check the use of the raw data for consistency with that of the provided analysis. Therefore, the data suppliers will be expected to be able to provide raw or basic data to users who wish to analyze it. If discrepancies exist the user should directly contact the data analyst at the site to discuss the differences and to inform the central facility.

Derived parameters, in the case of model outputs, should be stored as deemed appropri-
ate by the Modelling community and the contributors. Standard outputs of many models that require large amounts of computer time to generate should be available. In addition the input parameters for these submitted model runs should be available to the user.

It is very important that a centralized CEDAR data base facility be easy to access. A program like CEDAR which is attacking global aeronomical problems cannot succeed with measurements from only a few places. Foreign scientists in particular need assured access to data which can further their own studies. Access will be directly available through networking and less directly through magnetic tapes or other media. With the present incoherent scatter data base, most of the access is through magnetic tapes which are copied at NCAR and sent throughout the world. The common data format makes it much easier for scientists to use data from several sites. There is also on-site access for the scientists at NCAR and for short or long-term visitors, some of which are funded through the data base. The basic parameters from 6 incoherent scatter radars for a one day experiment consists of about 14 Mbytes of information \((1 \times 10^8 \text{ bits})\) in packed binary form. A worst case expansion to ASCII characters would increase this figure by a factor of 3. This is a great deal of data, which implies that if a user is interested in all the data from all the instruments for one campaign, which is often several days in length, it is much better to get copies of the data sent on magnetic tape or other media rather than try to analyze it, look at it or copy it over the network. With such large amounts of data, it is probable that plotting at the central facility will only be possible for summary data or for data that is not very dense. Copying data to tape can be at any level and copies over the network will probably be limited if the network time becomes excessive.

Most universities and research institutions in the contiguous 48 states have access to a computer network. Since most domestic CEDAR researchers are supported to some extent by the National Science Foundation, NSFnet should be available. A SPAN connection should also be accessible as should other INTERnet networks and BITnet. Copying, plotting, analysis, and comparisons with various models will be available over the network. Hopefully, the software written at the central facility and at other facilities will be portable enough that various software routines can be copied and used at a user's own site.
Extensive copying of data should be on magnetic tape or other media. Exabyte tape drives and controllers cost about $7500 and can store 2 Gb of data on an 8 mm cartridge, which can be re-written. The cartridges cost about $15. It is possible that data will be available on this medium in the future. CD-ROMs are another possibility which are attractive for users since readers are only about $1000. However, it costs about $8000 to make a CD-ROM master, which means they are only inexpensive if large amounts of disks are to be distributed. Since methods of deriving parameters differ, the CEDAR data is not as stable as geomagnetic data. In addition, some data suppliers may be very uncomfortable with the idea of their data distributed in mass quantities. Applications that need rapid access to large data sets may benefit from write once-read many (WORM) optical disk drives. Present prices are about $500 per 12 inch optical disk and about $15,000 for the drives. These prices are expected to drop in the future. Smaller optical disks are also a possibility. The CEDAR Data Base Committee will study these and other alternative media to 9 track magnetic tapes that can provide users with data in a convenient form.

To make most efficient and productive use of the information in the data base, the system must be interactive and easy to use. Current “desk-top” metaphor operating systems, such as Macintosh OS and Microsoft Windows, allow information to be quickly and easily accessed without the user being exposed to the underlying operating system, or even the keyboard through the use of a mouse or other pointing devices.

In addition to a “desk-top” environment to manipulate the data, the easiest way to assimilate large amounts of data is in a graphic format. To avoid confusion over different types of graphics systems, three basics types of graphics systems can be defined. In a batch graphics system, the user runs a program that produces an output that must be printed on an external device. If the plot needs alterations, the job must be re-submitted. A menu-driven graphics system allows the user, after running a program interactively, to see the plot on a display screen. However to change the plot requires re-running the program. An interactive graphics system allows the user to manipulate the plot on the screen. For instance the axis scale or labels may be altered through use of the keyboard or a pointing device, and the changes are immediately shown on the display screen. The
CEDAR Data Base system should support both black and white and color interactive graphics. These plotting packages should be easily accessed, either at the central location or using intelligent workstations at the user's home institution. Initially, however, the plots will be available through batch or menu-driven programs.

One potentially useful software package is the Image Reduction and Analysis Facility, IRAF (Tody, 1986). IRAF was designed by the National Optical Astronomy Observatories (NOAO) to provide a general purpose system for scientific data analysis. In addition to a variety of packages for image processing, image reduction and graphics, the system provides a full programming environment to allow users to implement their own analysis tools. The design of the IRAF system emphasizes portability between operating systems and computers.

The CEDAR Data Base Committee will study the possibility of using commercial data base management systems, as well as software developed by CEDAR participants such as the group at Millstone Hill Observatory. Electronic browse capabilities can be enhanced with commercially available artificial intelligence techniques, which allow the data base variables to be queried using English sentences or a pointing device rather than obscure computer syntax. One example of an application of a commercial package is the use of HyperCard (Goodman [1987]) for summary level browsing and manipulation of data. An inexpensive HyperCard system is currently available for Apple's MacIntosh computers, and is in the process of being written for DOS systems. HyperCard is a free-form data base that allows the integration of text, graphics and sound. The use of HyperCard will allow users to search diverse forms of data (including images) based on their own intuition, rather than requiring keyworks or field to be assigned in advance. The CEDAR data base should evolve to the point where a scientist does not need to understand the underlying computer architecture to access the data base information.

The focus of the data base to this point has been on the organization of information from NSF-sponsored CEDAR projects. To more fully utilize available information, relevant data from other government agencies such as NASA, NOAA, and DOD must be available. Networking or sharing data with other data systems, for example NASA machines dedicated
to future satellite projects such as UARS, will increase the scientific return of both the CEDAR and NASA projects. Many data sets of interest to the aeronomy community, such as geomagnetic indices, interplanetary magnetic field data and meteorological data, are now available on CD-ROM (optical disks similar to audio compact disks). For example, J. Allen of the World Data Center has made available a CD-ROM of geomagnetic data. Other organizations are following this example. "Jukebox" style changers for CD-ROM are now available which will allow users to remotely mount large data sets on optical disk to support their analysis efforts.

In addition to communication with other agencies involved in aeronomical research, increased communications within the CEDAR community are needed. There is a strong community consensus that electronic mail is an important part of the data base, to allow prompt communications to the data base staff as well as to other users and contributors. Much of the electronic communication among aeronomers utilizes NASA's Space Physics Analysis Network (SPAN), which currently only supports DECNET. This capability should be expanded to include the use of NSFnet to connect the centralized data base, individual users and the NSF Division of Atmospheric Sciences for increased (and less expensive) transfer of information. One of the files that will be available at the central facility and also distributed in print among the community will be a list of names, addresses, telephone numbers, and computer addresses of community members.

Preliminary reports from the CEDAR Data Base Group Leaders appear in Appendix B. These reports reflect initial discussion among the Group Leaders and scientists in their area. Many questions on the type, self-consistency, quality of data and documentation supplied to the data base need to be addressed immediately by these groups, as many issues are relevant only to specific instruments.

4. ORGANIZATION OF THE CEDAR DATA BASE

Successful operation of the database will require active participation of all concerned parties: the Data Base Committee, the Central Facility, data suppliers, and data users. In order to clarify the roles of each party, the committee proposes the following sets of
4.1. Responsibilities of the CEDAR Data Base Committee

The CEDAR Data Base Committee is responsible for organizing community inputs on development of a data organization structure that can best serve an international community of independent research groups. Specifically, the CEDAR Data Base Committee is responsible for:

1. Deciding what data types generated from an instrument or model should be submitted to the data base.

2. Coordinating with data suppliers from each discipline in cross-calibration of instrumental, modelling and analysis techniques to insure self-consistency in the basic and derived parameters submitted to the Data Base.

3. Coordinating with the CEDAR data base staff an acceptable format for submission of the data and to develop a timetable for data submission.

4. Establishing a dialogue in each research area of the CEDAR program to resolve data base issues related to the submission of data to the CEDAR Data Base.

5. Defining a user interface for CEDAR data base machines which simplifies accessing the system.

6. Defining what types of software and services are necessary for a useful data base.

7. Monitoring progress made at the central data facility towards the above goals.

The CEDAR Data Base Committee chairman is responsible for:

1. Participating in and providing inputs to the Data Base Committee.

2. Working closely with the data base staff to insure that the community goals defined by the Committee are being actively pursued.

3. Reporting directly to the CEDAR Steering Committee the current status and plans of the data base and related issues.
(4) Appointing CEDAR representatives to attend data base meetings of related agencies that interact with the CEDAR data base.

It is important that the Data Base Committee, particularly the Group Leaders, begin discussions of these issues with their colleagues as soon as possible. Some Group Leaders have contributed results of initial discussions, which appear in Appendix B. The Fall AGU meeting may be a good time for these sections to meet so data will soon be available to start the data base. In addition, the Data Base Committee suggests that an entire day at the next CEDAR meeting be devoted to data base issues.

4.2. Responsibilities of the Central Facility

The responsibilities of the central facility in supporting the data base are:

(1) Maintaining a catalog, ultimately to be available online, of contributions and standards for data in the data base.

(2) Archiving all relevant data sets and associated documentation.

(3) Coordinating data transfers between other agencies for ancillary data that supports the data base.

(4) Maintaining and managing an information system with adequate capabilities for users to browse, plot, and transfer data. Submission of data will be treated separately so it can be checked and controlled.

(5) Providing an up-to-date list of electronic mail addresses of the community and facilities to access SPAN, INTERnet (which includes NSFnet, ARPA.net and CSnet among others) and BITnet.

(6) Providing a data base staff to help with the submission and accession of data.

(7) Keeping informed of technical advances that may be of benefit to the data base.

(8) Planning data base additions and expansions as necessary.

(9) Cooperating with other institutions who may keep part of the total data base.

(10) Providing some travel funds for users to visit the central facility to access the data base system.
THE FIRST CEDAR DATA BASE REPORT

(11) Assuring that users have agreed to follow the "Rules of the Road" for use of the data base (these rules are discussed in a following section). Data suppliers will be given a copy of data requests for their data so they can be aware of the research going on involving them.

4.3. Responsibilities of the Data Supplier

The responsibilities of the data supplier includes:

(1) Cooperating with the appropriate Data Base Group to determine what data shall be submitted to the central facility and to determine an appropriate format that is sufficiently global in nature, since a common format for most types of data is desirable.

(2) Contributing adequately documented data (and catalogs of data) which satisfy criteria established by the appropriate Data Base Groups and contributing this data and documentation in a timely manner in an agreed upon format.

(3) Making raw and basic data available (possibly in an agreed upon format) to interested users along with calibration data.

(4) Contributing analysis software for creating derived parameters.

(5) Agreeing to consult with data users about the limitations and appropriate uses of the data.

4.4. Responsibilities of the User

The responsibilities of the data base user include:

(1) Strictly following the "Rules of the Road" and not using data or software packages in any presentations, reports, papers or other formats without an offer of co-authorship to the appropriate people.

(2) Acknowledging the CEDAR data base and other organizations as necessary in all publications.
(3) Notifying the data base of all uses of the data base in publications so an up-to-date list can be maintained at the central data base facility.

5. RECOMMENDATIONS FOR IMPLEMENTING THE CEDAR DATA BASE

5.1. Advantages of merging with the Incoherent-Scatter Radar Data Base

It is the general opinion of the CEDAR community that the most desirable location for the centralized data base facility is at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. NCAR, a NSF facility, is the home of the Incoherent-Scatter Radar Data Base (Wickwar [1984]). Since incoherent-scatter measurements of the ionosphere are critical for many CEDAR programs, it is clearly advantageous to merge the Incoherent-Scatter Data Base with the CEDAR Data Base at NCAR.

One of the most important parts of a successful scientific data base is the availability of knowledgeable scientists closely involved with the evolution of the data base. The data base scientists facilitate the contact between individual users and suppliers of data. By being active, interested users of the data, these scientists can help encourage the submission of clean data in a timely manner, and can help ensure that adequate software exists to allow users to access this information. The NCAR group already has experience in handling large and diverse aeronomic data sets and models, and it would greatly benefit the CEDAR program to use their expertise in establishing the CEDAR Data Base.

5.2. Data formats

The incoherent-scatter radar community has developed a tape format standard for Import/Export of data (Wickwar [1984]). The CEDAR community is urged to adopt this format for several reasons:

(1) It is flexible. It permits any type of data (including model output) to be ordered in any fashion, with any desired spatial and temporal resolution. This feature makes it easy for all groups to organize their data in a way most appropriate for their instruments.
(2) It is self-documenting. Each logical record fully defines all of the variables contained within it. This feature helps greatly to avoid confusion when working with different data sets, and makes it easy to keep track of the precise definitions of the data contained in the data set.

(3) It is compact. By using 16-bit integers rather than ASCII characters, several times more data can be stored and transmitted on any given medium. Although storage and transmission media have been rapidly increasing in capacity, the amounts of data taken and manipulated have been increasing at a comparable rate, so compactness will remain a desirable feature. In addition, using 16-bit integers is easier to implement than floating point numbers, which can vary in length and organization on different machines. There is some feeling that 32-bit integers should be used instead of 16-bit integers since there are so many 32-bit word machines. This change could be made fairly easily.

(4) It ensures data integrity for each block via checksums and word counts.

(5) It is already in wide use by a substantial fraction of the CEDAR community, and it is supported by NCAR.

An example of the NCAR tape format is shown in Appendix C.

The NCAR tape form has been successfully used by the incoherent-scatter radar groups. Other data formats have been developed and are in use by the space science community. One such format is NASA's Common Data Format (CDF; Treinish and Gough [1987]). The CDF is a highly flexible data format that can handled diverse data sets. It is presently available for Digital computers using the VAX/VMS operating system.

NASA has also developed the Standard Format Data Unit (SFDU) as a general purpose specification for data exchange (Space Station Data Network Concept, April, 1985). The SFDU is intended for any application where complete and formal data specification is required to facilitate automated operations, such as inter-process message exchange or data base management. SFDU consists of a primary label, which contains globally interpretable identification of the source and structure of the data set, and supplementary labels which
provide standard methods for describing the unique attributes of the encapsulated data. In the future the SFDU will be supported by a standard data definition language. The Data Base Committee will investigate whether these formats offer advantages for use by the CEDAR program.

5.3. **Data Base Computer**

The CEDAR Data Base Committee recommends that the central facility obtain a minicomputer and begin the process of establishing interactive access to the data base. The machine should use the UNIX operating system.

The choice of UNIX has several advantages. UNIX is an operating system standard, in use worldwide. It is primarily written in C and is thus highly portable and does not require proprietary hardware. Therefore, the operating system always looks the same to the user despite differences in the underlying hardware. Consequently, a large number of commercial and public domain software packages are available. NCAR has made a commitment to use UNIX for all its support computers, which means the operating system will be supported by the staff of the central data facility. Furthermore, the CODMAC [1986] report also urges the adoption of UNIX as a standard operating system for the space sciences. The use of UNIX at the central facility will facilitate the installation of the data base software that Millstone Hill Observatory is planning to port to UNIX.

The use of the UNIX operating system will allow distribution of the data base to be simplified, as users can choose from a wide variety of UNIX machines rather than being locked into a proprietary operating system. Because UNIX is so widely used, many emulators are available to simulate popular operating systems such as Digital's VAX/VMS operating system. "Desktop" environments are currently being developed for UNIX systems. The Inter-University Consortium for Education Computing is developing a windowing system for UNIX (Crecine [1986]). Since UNIX runs on everything from inexpensive workstations to mainframes (the operating system of the CRAY-2 supercomputer is an enhanced version of UNIX) the CEDAR community will benefit from using a widely supported operating system.
Networking of the CEDAR data base machine is extremely important. Initial plans are to have the machine connected directly to NSFnet. The SPAN connection would be converted to TCP/IP within the local NCAR network. ARPAnet and TELEnet are also available through network connections at NCAR. NSFnet is now in its infancy, but for a NSF supported program like CEDAR it may provide the easiest access to the data base for the domestic community.

The data base staff should attempt to provide the software necessary to copy the data, sort it according to user-specified criteria, and plot it. Portability will be important and many of the plotting packages and software used should either be free or available for a nominal charge. The Millstone Hill data base system should also be incorporated into the central facility. The central facility and Millstone Hill Observatory should work closely with each other to share expertise as well as software. Advice and expertise is also available from the larger group of computing personnel on site at NCAR for various limited purposes.

The initial CEDAR data base implementation should use existing data at NCAR, which is predominately from incoherent-scatter radars. In meetings and discussions over the next several months the Group Leaders should decide, with the help of researchers in their field, timetables for submission of high-quality, consistent data to NCAR. Before attempting to process old data sets, the highest priority should be given to data from current CEDAR campaigns and ongoing programs.

5.4. “Rules of the Road” and Security Concerns

The combined experience of several NASA satellite programs and the initial operations of the Incoherent-Scatter Radar Data Base have produced a set of conduct guidelines for data base use. In Appendix D “Rules of the Road” for the CEDAR Data Base are suggested. With minor changes, these are the same guidelines presently used by the Incoherent-Scatter Radar Data Base. These rules should be considered for adoption for use by the CEDAR Data Base. If the users want changes or modifications to these rules, concerns should be forwarded to the CEDAR Data Base Committee. For successful operation of the data base it is important that these rules be followed.
A topic open for community discussion is that of data base security. Groups contributing data to CEDAR campaigns may want to use the structure and resources of the data base but may feel that their campaign results are proprietary and access should be restricted to certain individuals or groups. Do limitations need to be put on access to the data base? Should non-CEDAR participants have the same access rights as others, providing they follow the "Rules of the Road?" The current policies of the NCAR Data Base allow any interested scientist who agrees to the "Rules of the Road" to have access to all data, and in general it is recommended that the same policies apply to users of other CEDAR data. However, these important issues need community-wide discussion at the next CEDAR meeting.

5.5. *The Bottom Line*

A wide range of issues concerning the structure and purpose of the CEDAR Data Base have been discussed in this report. Implicit in this discussion is that adequate support from the NSF will be available to implement this important part of the CEDAR program. Not only does the central data facility need sufficient funding to support data base operation, but funding for prompt and proper data analysis and computer simulations must be made available to the CEDAR community and the data base. The NSF must show its support for CEDAR projects as *long-term community efforts*, not short-sighted endeavors. Many of these programs will require acquisition of data over extended periods. These data must be properly documented and preserved in an accessible manner, so future scientists may use them to further our understanding of the coupling, energetics and dynamics of the atmosphere.

6. **LONG RANGE OBJECTIVES OF THE CEDAR DATA BASE**

The plan for a centralized data base system outlined in the previous sections is a necessary beginning for data management in the CEDAR program. However, the community needs to be aware of developing ideas in data base systems and be prepared to follow new directions that will increase the scientific output of the CEDAR program.
THE FIRST CEDAR DATA BASE REPORT

Many research groups are conducting experiments in distributed data base techniques. One such experiment is the use of the Browse System, a University of California at Santa Barbara project, to remotely access and analyse remote sensing data. The University of California at Berkeley’s Extreme Ultraviolet Explorer (EUVE) project is evaluating this system for incorporation into its guest observer program. The EUVE guest observers will have access to imaging photometer and spectrometer data from the spacecraft. A preliminary report is included in Appendix E.

Briefly, this work is an example of telescience, where both the operation, data analysis and data base functions for a given satellite (or remote ground-based instrument) are controlled over a computer network, allowing many remotely situated experimenters to both control the experiment and work with the data. Class I CEDAR observatories will require remote control and analysis of individual experiments. Such control techniques will ultimately make the observatories less expensive to operate by reducing the amount of observer time required at remote locations.

Artificial intelligence techniques should also be considered to control instruments in a Class I observatory. Many experimenters have automated their instruments so the machines can turn themselves on and off at specified times, perform calibrations, etc. However, with currently emerging artificial intelligence techniques, the instruments could incorporate more advanced functions, such as modification of observing programs due to clouds or geomagnetic activity. The CEDAR program should encourage further study of these methods, in addition to the use of artificial intelligence to improve browsing capabilities of the data base as previously discussed.

Acknowledgments. We would like to thank the many members of the CEDAR community who have already contributed to this report. The National Center for Atmospheric Research is sponsored by the National Science Foundation.
Committee on Data Management and Computation (CODMAC), Issues and recommendations associated with distributed computation and data management systems for the space sciences, National Academy Press, Washington, D. C., 1986.


Ground-Based Optical Aeronomy Science Steering Committee (GBOA SSC), Coupling, Energetics, and Dynamics of Atmospheric Regions (CEDAR), Volume 1: Overview, April, 1986.


# The First CEDAR Data Base Report

## Table 1

### Four Levels of Radar Data

<table>
<thead>
<tr>
<th>Raw Data</th>
<th>Basic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autocorrelation Functions or Spectra</td>
<td>Corrected Autocorrelation Functions or Spectra</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power Measurements</th>
<th>Basic Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute Power Profiles</td>
<td>Electron Temperature</td>
</tr>
<tr>
<td></td>
<td>Ion Temperature</td>
</tr>
<tr>
<td></td>
<td>Ion Composition</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Derived Parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Ion Velocity</td>
<td>Electric Field</td>
</tr>
<tr>
<td>Meridional Neutral Wind</td>
<td>Vector Neutral Wind</td>
</tr>
<tr>
<td>Exospheric Temperature</td>
<td>Neutral Temperature</td>
</tr>
<tr>
<td>Atomic Oxygen Density</td>
<td>Heat Flux at High Altitude</td>
</tr>
<tr>
<td>Energy loss from Electrons to Ions and Neutrals</td>
<td>Energy Input from Thermal Conduction</td>
</tr>
<tr>
<td>Hall Conductivity</td>
<td>Pedersen Conductivity</td>
</tr>
<tr>
<td>Current Perpendicular to $\vec{B}$</td>
<td>Birkeland Current</td>
</tr>
<tr>
<td>Joule Heating</td>
<td>Ion-Electron Production Rate</td>
</tr>
<tr>
<td>Energy Deposition by Auroral Electrons</td>
<td>Energy Spectrum of Auroral Electrons</td>
</tr>
<tr>
<td>Several Optical Emissions</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Anticipated growth rates for NASA Solar and Space Physics data, from the CODMAC [1986] report.
THE FIRST CEDAR DATA BASE REPORT

APPENDIX A

THE CEDAR DATA BASE COMMITTEE

S. Chakrabarti, University of California-Berkeley, Spectroscopy Group Co-Leader
C. Deehr, University of Alaska, Spectroscopy Group Co-Leader
B. Emery, NCAR
C. Fesen, University of Colorado, Modelling Group Leader
J. Forbes, Boston University, Middle-Atmospheric Radar Group Leader
C. Gardner, University of Illinois, Lidar Group Leader
M. Harel, Jet Propulsion Laboratory
J. Holt, Massachusetts Institute of Technology Haystack Observatory
T. Killeen, University of Michigan, CEDAR Steering Committee
G. McCormac, University of Michigan
M. Mendillo, Boston University, Imaging Group Leader
A. Richmond, NCAR, Incoherent-Scatter Radar Data Base
W. Sharp, University of Michigan, CEDAR Steering Committee
R. Sica, Utah State University, Chairman
C. Tepley, Arecibo Observatory, Interferometry Group Leader
V. Wickwar, SRI International, Incoherent-Scatter Radar Group Leader
Incoherent-Scatter Radars
V. B. Wickwar

In 1981, the incoherent-scatter radar community started to examine the possibility of creating a data base. Detailed discussions occurred at a workshop, at NCAR, in February 1982 and the incoherent-scatter data base was established at NCAR in 1983. Because it has proved very useful to a large segment of the aeronomy community, a brief review of the data base is provided along with suggested extensions of it that should be incorporated in the CEDAR data base.

At the heart of the data base is a tape format for exchanging data. It is flexible, yet well defined. It was intended primarily for the exchange of basic parameters derived from the radar measurements, such as electron density, electron and ion temperatures, and ion velocity, as well as a few parameters derived from these, such as vector velocities. It has been used for many other derived parameters such as conductivities, meridional neutral wind, vector neutral wind in the lower thermosphere, joule heating, and particle energy deposition. It has also proved very useful for IMF parameters from IMP 8, particle fluxes from the NOAA satellites, optical intensities, and outputs from the NCAR TGCM.

Using this format since 1984, a considerable amount of data, 7.5 Gbytes, has been submitted to the data base from seven incoherent-scatter radars: Sondrestrom, Chatanika, Millstone Hill, Arecibo, Jicamarca, EISCAT, and St. Santin. Initially the data were submitted starting with 1983. The data are now submitted two to nine months after the observations. Recently, data going back as early as 1966 have been submitted from several radars and early data are expected from additional ones in the future. A full list of available data is in the data base catalog that can be obtained from Barbara Emery at NCAR (HAO, PO Box 3000, Boulder, CO 80307) or can be accessed on the incoherent-scatter data base login on the NCAR IBM 4381 computer.

Data can be obtained on tape from the data base using the same format by individuals who have agreed to follow the "Rules of the Road." As intended, easy access to the data base has led to new users of the radar data and to new scientific studies. Data and software have been accessed by about 40 scientists in the U.S., Europe, and Asia. New studies include global-scale investigations of ion convection and neutral winds, and solar-cycle variations.

In addition to these benefits, several other less obvious ones have occurred. There is now more standardization in data reduction assumptions and procedures as well. This arises because of the greater ease of comparing simultaneous data from several radars. The
data reduction is more systematic and more timely. In addition, the fact that all data are in a common format has made software development easier for display, correlation studies, and statistical analysis. Most importantly, it has become much easier to combine data from several instruments, radar and non-radar, and to combine data with the results of theoretical simulations. This last benefit is, of course, at the very heart of the CEDAR initiative.

In the future, the CEDAR data base should extend these many benefits to the other data produced and needed by the aeronomy community. In addition to the basic exchange of data, some new features would benefit the data base. One, which a part of the radar community has used and liked, is having the data accessible on line, as is possible with the data base at Millstone Hill. Another is having an extensive browse capability as is now possible with the collections of hard copy.

In the future, the CEDAR data base should extend these many benefits to the other data produced and needed by the aeronomy community. In addition to the basic exchange of data, some new features would benefit the data base. One, which a part of the radar community has used and liked, is having the data accessible on line, as is possible with the data base at Millstone Hill. Another is having an extensive browse capability as is now possible with the collections of hard copy color pictures and plots produced for Sondrestrom.
The following is a description of the data requirements for spectroscopy in the CEDAR program. Note that these requirements in no way represent the consensus of the whole community, rather they are parameters of interest for most spectroscopic experiments. We have listed the parameters in the same format as shown in Table 1 of this report.

The table does not address any of the data collection, operation and distribution issues of the Data Base. We have also excluded technological issues such as modes of storage (video tape vs. WORM), graphics support, etc.

CEDAR SPECTROSCOPY DATA REQUIREMENT

(1) RAW DATA
   (a) wavelength intensity
       1. intensity (raw bits/s)
       2. wavelength (nm)
   (b) calibration procedure
       1. source (°K)
       2. source distance (m)
       3. source radiance (Wm⁻²)
       4. incidence angle (degrees)
   (c) instrument type
       1. aperture (mm)
       2. collimator focal length (mm)
       3. camera focal length (mm)
       4. grating blaze angle (degrees)
       5. grating line density (lines/mm)
       6. grating size (length (mm) x width (mm))
   (d) optics parameters
       1. order (n)
       2. prefilter (type, or Corning or Schott no.)
       3. integration period, frame (s)
       4. wavelength bin size (nm)
       5. slit size (length (mm) x width (mm))
       6. short wavelength (nm)
       7. long wavelength (nm)
       8. no. of integration periods
   (e) detector parameters
       1. no. of pixels (n x m)
       2. pixel size
THE FIRST CEDAR DATA BASE REPORT

3. operating temperature
4. dark response vs. pixel
5. quantum efficiency vs. pixel
6. readout frequency

(2) BASIC DATA
(a) instrument
   1. field-of-view (steradians)
   2. area (cm²)
(b) wavelength emission rate
   1. emission rate (Rayleighs/nm)
   2. wavelength (nm)
   3. resolution (nm/nm)
(c) Universal Time of exposure
   1. date (yyyy:mm:dd)
   2. hour (hh:mm:ss:frame no.)
   3. duration (hh:mm:ss:frame no.)
(d) geographic position
   1. latitude (decimal degrees)
   2. longitude (decimal degrees)
(e) direction
   1. altitude (decimal degrees above local horizon)
   2. azimuth (decimal degrees from N through E)
(f) weather
   1. sky cover (tenths)
   2. extinction (%)

(3) BASIC PARAMETERS
(a) peak wavelength emission rate
   1. emission rate (Rayleighs)
   2. wavelength at peak (nm)
(b) wavelength integrated emission rates
   1. emission rate (Rayleighs)
   2. low wavelength (nm)
   3. high wavelength (nm)
(c) emission ratio
   1. wavelength of numerator (nm)
   2. wavelength of denominator (nm)
   3. peak or integrated
   4. value of ratio
   5. error of ratio
(4) DERIVED PARAMETERS

(a) synthetic spectral fit program
1. program source
2. goodness of fit (%)
3. Doppler temp. (°K)
4. rotational temp. (°K)
5. vibrational temp. (°K)
6. species

(b) model atmosphere
1. source
2. month
3. exosphere temperature (°K)
4. magnetic indices
5. 10.7 cm flux

(c) column distribution of emission
1. beginning distance from observer (m)
2. ending distance from observer (m)
3. functional dependence (m)

(d) incoming particle type
1. electrons (y/n)
2. protons (y/n)
3. other ions (y/n)

(e) incoming particle
1. flux (number cm$^{-2}$ s$^{-1}$ ster$^{-1}$)
2. energy (ergs cm$^{-2}$ s$^{-1}$)
3. average energy (ev)
4. characteristic energy (ev)
5. efficiency (Rayleighs/erg)

(f) fourier analysis
A. H. Manson, Chairman Lower Thermospheric Coupling Study (LTCS) Radar Committee

A global network of radars, many of which played vital individual roles in MAP/MAC, has now been organized for participation in WITS and CEDAR projects including LTCS, GITCAD and WAGS. There are over 20 mesosphere and lower-thermosphere (MLT) systems, including 9 medium frequency (MF) radars (formally called partial reflection radars), 3 MST radars, and over 8 meteor radars (VHF) since the GLOBNET system is included.

Data are being collected for the first LTCS campaign (1 month), and a master tape prepared at ISAS, University of Saskatchewan (Drs. Manson and Meek), for distribution to all radars and WITS/CEDAR data bases. LTCS approved scientific projects have first use of these data which include mean winds, and amplitudes and phases of the 2-D, 24-H, 12-H, 8-H oscillations, with gravity wave intensities in two spectral bands (four day means are used). Full data format descriptions for MLT radars will accompany the tape.
Contributions of Optical Interferometers to the CEDAR Database

It is important to establish a consensus among those of the aeronomy community involved in optical observations about the type of data that would form a useful part of a CEDAR database. My responsibility was to survey the opinions of those who use optical interferometers in atmospheric studies. A questionnaire was distributed in July, 1987 to 25 colleagues who work with Fabry-Perot or Michelson interferometers, both in the U.S. and abroad. Only 12 people returned the questionnaire, which is a rather disappointing response. The following is a summary of the comments that I received.

In the survey, we first wanted to get an idea of what instruments were available, which of them could potentially contribute to an optical database, and which emission lines were routinely measured. Thirteen observing sites were identified by our survey, although we know of at least six others that are operational. There are also five sites whose operational status is uncertain. All those who answered said they would try to contribute both old and newly acquired data, where possible, and everyone expressed an interest in using such a database in their research.

Another section of the questionnaire was concerned with the type of data that should be archived, how it should be formatted and managed for efficient transfer to a database, and what special individual circumstances exist, if any, that might either impede or be of benefit to the establishment of the database. With regard to data formats, the survey asked for the type of machine that was used for routine data reduction. We also asked whether respondents had any experience using the NCAR Incoherent Scatter Radar Database or other archives of atmospheric data.

The majority of the workers observe the emissions from either O('D) or O('S) at 630 and 557 nm, respectively, with the red line being the most popular. Systematic studies of H\alpha, OH, and O+ emissions are also in progress at a few sites. Data obtained from these instruments can be classified into two groups, 1) basic data and, 2) derived parameters. The difference between the two is that few assumptions are made when basic data are derived from instrument dependent information, like count rate versus pressure or voltage. For certain derived parameters, model assumptions are sometimes required, for example, neglecting spatial gradients when determining the meridional or zonal wind components. For basic data, those parameters identified by respondents as constituting a significant and minimal part of the archives were:
The date, time, azimuth and elevation angles of the observation, integration period of the measurement, and geographic and magnetic coordinates of the station. Other relevant data could be added here, such as the weather conditions during the time of the observation, if known, plus the assumptions made in the data reduction. These last points might be described in a header that accompanies the data records.

Line-of-sight velocity and error derived from the Doppler shift of the emission. The method of removing the baseline or 'zero' Doppler information should be defined since there are several ways that this might be accomplished.

Temperature and error determined from Doppler widths. Again, it is important to state the data reduction process. For example, is the instrument function properly deconvolved from the measured emission shape or is a Gaussian lineshape assumed?

Emission intensity and possibly background radiance and their errors. These data should be normalized to a photometric measurement, if possible. Otherwise, a relative value might be stored provided that it is flagged properly.

Examples of derived parameters that were thought by eight respondents to be important input to a database were:

- Meridional, zonal, and vertical wind velocity for the 'overhead' location of the observing station.

- Horizontal gradients in the wind and temperature fields.

- Derived densities of the emitting species. This may not be too complicated as in the case of a resonant fluorescence emission like \( \text{H}_\alpha \), or when independent supplemental information, for example from a radar, may be available.

- Velocity distribution derived from non-thermalized or non-Maxwellian emission profiles. Such input to the database would necessarily include the raw data on which different analysis methods might be applied.
Optical interferometry can make a significant contribution to a CEDAR database if everyone attempts to provide at least the minimal basic data set to those archives. Derived parameters require more work, of course, but can enhance the overall quality of the database. It would be left to the individual to decide what information to provide beyond the basic set. It would also be the responsibility of the data provider to assure that his/her data are of good quality and consistent with the rest of the community. As an obvious example, the data should be free of contaminating effects, such as clouds or other noise sources - or at least be flagged accordingly.

Concerning the volume of data generated from optical interferometers, with the exception of the field-widened instrument with an array detector, most interferometers do not generate a large amount of data. For these devices alone, data compression is unnecessary. However, the storage of data should probably have a universal format for all types of optical instruments in order to facilitate ease of intercomparison of their results. For the high volume output instruments such as imaging photometers, data compression is needed. It would make good sense for the interferometer community to adopt a data format that satisfies the needs of all.

A couple of respondents expressed a concern on this matter. Where local databases exist, there is some justifiable resistance to retool their efforts in order to satisfy the input requirements for the archives. Generating a reformatting program is not a trivial task. For these situations, we suggest that the holders of the archives, whether it is at NCAR or another facility, provide assistance and reformatting programs to ease the workload of the data provider. The local database structure can be maintained as input to a 'black box' that converts it to archive data in the required format. The majority of those who answered the questionnaire did not feel that reformatting was a formidable task. Nearly all of us use a VAX computer to reduce data. This makes the reformatting question even less important since we have near-universality already.

Finally, one key issue that we all should be concerned with is data integrity. No one wants to use a database that contains shaky or questionable results. A few of the respondents expressed an interest in establishing a set of standards among the various methods used for analysis of the interferograms. We would have to agree on the best method by which to reduce our data.

For winds, how do we determine the rest frame Doppler? The ideal situation without applying any assumptions would be to compare the measured sky fringes with a lamp emitting at the same
wavelength, like the oxygen source described by Peterson et al. [Planet. Space Sci., 27, 1209, 1979]. But, such systems are not readily available and have a short lifetime. Alternatively, a simple difference between the measured north and south winds will yield the meridional component, and a similar operation can be done for the zonal winds. But, this usually assumes that the vertical component is constant across the span of the measurements, which is not always true. For temperatures, simple methods like Gaussian fitting will yield a reasonable result, but may not always be accurate. It is usually better to adopt a Fourier decomposition approach with the instrument function determined from a frequency-stablized laser or hollow cathode lamp.

Similar questions about data integrity came up among the radar community when the NCAR radar database was being established. There was agreement to adhere to a few standards, such as reducing the radar data using a nonlinear least squares fitting approach instead of various simpler methods. Only in this manner is consistency among the results assured. I hope the optical community will adopt a similar attitude to make the CEDAR database work.

Craig Tepley
Arecibo Observatory, October 1987
23 October 1987

TO: R. Sica
FROM: M. Mendillo, J. Baumgardner
RE: Considerations For An Imaging Component of the CEDAR/NCAR Data Base

1. INTRODUCTION

The two basic aspects of a common database involve, firstly, decisions on the scientific content of the observations to be shared and, secondly, the operational scheme by which collection and distribution are achieved. For geophysical imaging data, the scientific content plan would involve decisions on such factors as calibration methods, how background levels are determined, whether data should be corrected for instrumental effects such as lens vignetting and/or atmospheric Van Rijn effects, etc. In the preliminary discussion attempted here, we do not address the scientific context issue. Rather, we move directly to the operational issues of how to collect, archive and distribute "images", whatever that might mean.

2. OVERVIEW OF INPUT/OUTPUT FOR AN IMAGING DATA BASE

A. Original Data Sources are collected by a PI and stored in only two possible ways:

(1) Analog data:
- Hard copy
- Video tape
- Video disk
- Other (?)

(2) Digital data:
- 9 track tape
- WORM laser disk
- Other (?)

B. The Archived Data Base at some central facility would need to perform two basic tasks:

(1) Catalog Function -- tell potential user what is available.

(2) Data Storage Function -- assuming both A and D types of data, the two likely methods would be:

-- Video disk for ANALOG DATA
-- WORM laser disk for DIGITAL DATA
C. **User Access to Data Base** requires that the Central Facility is able to send images, as determined by (A) and (B) above, in a timely and inexpensive way.

The possibilities include:

(1) For Analog data,
- Video tapes via postal service
- Video disks via postal service
- Hard copies via postal service
- Browsing of potentially all images via "video-telephone-like" link.

(2) For Digital data,
- Floppy disks via postal service
- WORM laser disks via postal service
- CD-ROM via postal service
- Direct computer-to-computer transfer via data network(s), probably involving image-compression schemes.

D. **Issues** to be addressed include:

(1) What level of standardization should be imposed on PI-generated data to be contributed to the data base?

(2) What limitations should be set for possible formats/media to be made available to the data base user?

(3) Others (?)

E. **Suggested approaches to issues:**

(1) Maximize the possible ways of PI input methods to the data base. Standards are already in place for all of the possible original data sources (2 (a) above), except for WORM drives.

(2) Key output method to amount of data requested.

As an example, for requests of:
- (1 to $E$) digital image(s), use IBM floppy disks,
- ($E$ to $n$) digital images, use CD-ROM,
- (1 to $E'$) night(s) of analog data, use video tapes,
- ($E'$ to $n'$) nights of analog data, use video disks,

where $E$ and $E'$ are "a few" and $n$ and $n'$ are "a lot".
**Requirements for Various User Goals:**

<table>
<thead>
<tr>
<th>User Goal</th>
<th>Requirements at Central Facility</th>
<th>Requirements at User Site</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1) Slide or Prints</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a) sufficient quality for quick-look analysis</td>
<td>Standard Copy Facility -- Xerox (B&amp;W, Color)</td>
<td>None</td>
</tr>
<tr>
<td>b) Publishable quality</td>
<td>-- Photographic Capabilities -- Video Printer</td>
<td></td>
</tr>
<tr>
<td><strong>2) Video tape or video disk of specific night(s) or period(s)</strong></td>
<td>Video tape and video disk reproduction equipment</td>
<td>a) standard VCR or video disk player</td>
</tr>
<tr>
<td>a) view for quick-look analysis</td>
<td></td>
<td>b) PC/Image processing system to grab and process desired image(s)</td>
</tr>
<tr>
<td>b) work with digital image(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3) Access to digital data via:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a) 9 track tape</td>
<td>a) Ability to make 9-track tapes</td>
<td>(a) through (d) requires appropriate reading device and image processor --- ranging from a a PC-type ($3K) to a μ-Vax/SUN level system ($100K).</td>
</tr>
<tr>
<td>b) WORM disk</td>
<td>b) WORM laser disk drive</td>
<td></td>
</tr>
<tr>
<td>c) CD-ROM disk</td>
<td>c) CD-ROM master unit</td>
<td></td>
</tr>
<tr>
<td>d) Floppy disk</td>
<td>d) Floppy disk drive</td>
<td></td>
</tr>
<tr>
<td><strong>4) Real-time access of low resolution (compressed) images</strong></td>
<td>Computer resources (disk space, telecommunication, staff) to maintain on-line services</td>
<td>High speed modem or network system; output devices to view the desired images</td>
</tr>
</tbody>
</table>

**Issues:**

1. How removed should the PI responsible for the data be from the distribution system?
2. How different can "search the database" images be from the ultimate "publication quality" images?

**Suggested approaches:**

1. Have PI furnish data to Central Facility in easily-used, non publication quality format: VIDEO TAPE (example enclosed).
   -- The video tape would have images with grids and calibrations required for analysis.

2. Have user contact PI for special processing -- both scientific analysis and publication quality reproduction.
A wide variety of lidar techniques have been developed to study the chemistry and dynamics of the troposphere through lower thermosphere. The data base requirements can be substantially different depending on the lidar technique and the specific measurement. Because Rayleigh scatter lidars and sodium lidars are now widely used to study mesosphere dynamics, this initial report will concentrate on the data requirements for these two techniques. Requirements for differential absorption (DIAL), Raman and aerosol lidars will be developed in future reports.

The raw data for Rayleigh and sodium lidars consists of photon count profiles. These profiles are obtained by using a range-gated photon counter to measure the scattered signal versus altitude from each laser pulse. Because the count levels at the higher altitudes are very small, the laser is typically fired several hundred times and the photon counts from each shot are accumulated. For Rayleigh and sodium lidars, the photon count profile is proportional to density divided by the square of the range. Some form of gain switching is usually required to attenuate the low altitude signal to prevent overloading of the photo-multiplier tube (PMT). A typical sodium lidar photocount profile is plotted in Figure B.1. Gain switching can be accomplished by placing a rotating shutter in front of the PMT or by electronically changing the PMT sensitivity. Both techniques can introduce slight variations in the detector sensitivity after the gain has been switched to the maximum value. Consequently, it may be necessary to compensate the raw photocount profiles for these gain variations.
Figure B.1. Lidar photocount profile obtained on the night of January 20, 1987 at the Mauna Kea Observatory, Hawaii. The profile was collected by integrating the counts from 750 laser shots during a period of 100s starting at 2105 LST. The data were smoothed by computing a 750 m running average.
Density and neutral temperature are the basic data that can be computed from the photon count profiles. A wide variety of other geophysical parameters can derived from the density and temperature profiles [Chanin & Hauchecorne, 1981; Gardner and Voelz, 1987; Fricke & von Zahn, 1985]. Because the derived parameters are often calculated using models and other assumed parameters, the Rayleigh and sodium lidar data base will be restricted, initially, to photon count, density and temperature profiles. Table I is a list of the lidar system parameters required in the data base and Table II lists the lidar measurements in the data base.

References


TABLE I
SYSTEM PARAMETERS
RAYLEIGH AND SODIUM LIDAR DATA BASE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>(deg, min, sec)*</td>
</tr>
<tr>
<td>Longitude</td>
<td>(deg, min, sec)</td>
</tr>
<tr>
<td>Altitude MSL</td>
<td>(m)</td>
</tr>
<tr>
<td>Laser Beam zenith Angle</td>
<td>(deg)</td>
</tr>
<tr>
<td>Range Bin Length</td>
<td>(m)</td>
</tr>
<tr>
<td>Laser Pulse Length FWHM</td>
<td>(cm)</td>
</tr>
<tr>
<td>Wavelength</td>
<td>(pm)</td>
</tr>
<tr>
<td>Laser Linewidth</td>
<td>(pm)</td>
</tr>
<tr>
<td>Pulse Energy</td>
<td>(mJ)</td>
</tr>
<tr>
<td>Pulse Reception Rate</td>
<td>(pps)</td>
</tr>
<tr>
<td>Laser Beam Divergence</td>
<td>(μrad)</td>
</tr>
<tr>
<td>Receiver Aperture Area</td>
<td>(cm²)</td>
</tr>
<tr>
<td>Receiver Bandwidth FWHM</td>
<td>(pm)</td>
</tr>
<tr>
<td>Receiver Field of View FW</td>
<td>(μrad)</td>
</tr>
<tr>
<td>Overall Receiver Efficiency (including quantum efficiency)</td>
<td>(%)</td>
</tr>
<tr>
<td>PMT Quantum Efficiency</td>
<td>(%)</td>
</tr>
</tbody>
</table>

*Units denote required precision
TABLE II
RAYLEIGH MEASUREMENTS AND SODIUM LIDAR DATA BASE

<table>
<thead>
<tr>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective PMT Gain Profile</td>
<td></td>
</tr>
<tr>
<td>Raw Photocount Profile</td>
<td></td>
</tr>
<tr>
<td>Gain Compensated Photocount Profile</td>
<td></td>
</tr>
<tr>
<td>Density Profile (Sodium)</td>
<td></td>
</tr>
<tr>
<td>Effective backscatter cross-section ((cm^2))</td>
<td></td>
</tr>
<tr>
<td>Normalizing altitude ((m))</td>
<td></td>
</tr>
<tr>
<td>Assumed Pressure ((mb)) and Temperature ((K)) at normalizing altitude</td>
<td></td>
</tr>
<tr>
<td>Vertical resolution ((m))</td>
<td></td>
</tr>
<tr>
<td>Column Abundance ((cm^{-2}))</td>
<td></td>
</tr>
<tr>
<td>Centroid Height ((m))</td>
<td></td>
</tr>
<tr>
<td>RMS Thickness ((m))</td>
<td></td>
</tr>
<tr>
<td>RMS density error ((cm^{-3}))</td>
<td></td>
</tr>
<tr>
<td>Temperature (Rayleigh)</td>
<td></td>
</tr>
<tr>
<td>Assumed temperature at upper level ((K))</td>
<td></td>
</tr>
<tr>
<td>Vertical resolution ((m))</td>
<td></td>
</tr>
<tr>
<td>RMS error ((K))</td>
<td></td>
</tr>
<tr>
<td>Date UT ((yr, mo, day))</td>
<td></td>
</tr>
<tr>
<td>Profile start time UT ((hrs, min, sec))</td>
<td></td>
</tr>
<tr>
<td>Profile Integration Time ((sec))</td>
<td></td>
</tr>
</tbody>
</table>
1. Introduction

One of the motivations for the CEDAR program was the recognition that the development of global theoretical models requires the availability of large, coordinated observational data sets to validate the models. Further progress in understanding the upper atmosphere requires the adoption of a world view, i.e., the data analysis and theoretical modeling must begin to account for the coupling between atmospheric regions, both horizontally (e.g., between the high latitude and low latitude atmosphere) and vertically (e.g., between the mesosphere and thermosphere). Sporadic observations by isolated instruments must give way to coordinated campaigns, involving many instruments operating during the same period. The datasets obtained during these campaigns are most suitable for comparison with the theoretical models.

The analysis and interpretation of the observational data collected during the campaign periods may be facilitated by the availability of theoretical calculations for similar geophysical conditions. Conversely, theoretical model results may help guide the observations by providing predictions of the state of the atmosphere: e.g., what magnitude winds or exospheric temperatures are anticipated during geomagnetic disturbances. Therefore, it was decided to incorporate theoretical model calculations into the CEDAR database to assist in the acquisition and interpretation of data from the CEDAR campaigns.

In order to increase the utility of the modeling database, it will also contain ancillary models and software often required in theoretical and observational investigations. This includes, for example, empirical models such as the MSIS neutral atmosphere and Chiu ionosphere models, and subroutines such as the conversion from geographic to geomagnetic coordinates and representations of the geomagnetic field. The easy availability of these tools to the entire community will help ensure the close collaboration between the modelers and the observers which is vital to the goals of the CEDAR program.

2. The Modeling Database

- Contents

The contents of the modeling database may be very broadly categorized as follows, with examples of each category also listed:
- Theoretical/numerical model outputs: TGCM's, global ionospheres, dynamo model;
- Interactive codes: Strickland electron transport code; auroral spectra code;
- Auxiliary subroutines: magnetic coordinate conversions, magnetic field.

The acquisition and storage of the empirical models and interactive and auxiliary codes are fairly straightforward. The incorporation of the theoretical model results in the database requires greater effort. This is due to the size of the numerical model output, and to the necessity of documenting the geophysical conditions specific to the model run.

The theoretical model runs will be identified by a header record or records, which will fully describe the conditions and model inputs used to generate the solutions. This would include, for example, for model runs by the NCAR TGCM, the electron density model used to calculate the ion drag, the magnetospheric cross-tail potential, the 10.7 cm solar flux, and other parameters used or special conditions. The headers are extremely important, because only through them can outside users determine whether a particular model run is appropriate to the analysis of a particular data set.

Initially, the theoretical model runs will be confined to representing “average” conditions, i.e. sets of model calculations will be available for solar minimum or maximum, for each month of the year, and for quiet or disturbed conditions. These standardized runs will provide a medium for the experimenters to initiate contact with the modelers by providing a standard curve for data comparison. Modeling of individual days’ observations will be done at the modelers’ discretion, and will probably involve several runs.

- Storage and Formatting

Since the CEDAR campaigns are designed to be multi-instrument, multi-site observations of the upper atmosphere, the theoretical models must aim to provide descriptions of all observed quantities. These include, but are not limited to: neutral winds, temperatures, and densities, airglow emission intensities, particularly 5577 A, 6300 A, and the OH emissions, electron and ion densities and temperatures, plasma drifts, electric fields, and metallic ion densities. Model predictions can be provided over individual stations or globally for the 3-D models. Often, the model inputs must also be carefully recorded in the same manner. For the NCAR TGCM, model inputs would include electron densities, ion convection velocities, and auroral precipitation, among others.

Storage of the model inputs and outputs must be as efficient as possible, minimizing the required disk space. Model inputs and outputs will probably be stored on the NCAR
mass store in the export format used by the present Incoherent Scatter Data Base or in some other agreed upon common format. Some TGCM results have already been sent to users on tape in this format. A more condensed way to store the results may be in spherical harmonics such as Killeen's VSH (vector spherical harmonic) model has done. Decisions must be made concerning the acquisition of the model results: how and when will model runs be disposed to the database, and who is responsible for the collection of results, the modeler, or the database facility. It is anticipated that the experience gained through development of the NCAR Incoherent Scatter Database can provide some guidance. Accordingly, discussions are underway with those involved in that effort.

- Use

The database must provide two services to potential users:

1) provide a description of models and model results available;
2) provide the models and/or model results in a simple and usable manner.

The form of the model results will depend on the users' purposes. Digital listings and plots can be made from the mass store files. Network transfers and tape copies are also possible.

The modeling database will be freely accessible to the community. For the published, established models such as MSIS, IRI, SLIM, etc, there is no problem with this unrestricted access. For the theoretical model results such as those of the NCAR or UCL TGCM's, an "etiquette" must be formulated, since certain model runs may not be in the public domain.

The use of the modeling database may be guided by the rules formulated for use of the NCAR Incoherent Scatter Database. Briefly, this requires a user to contact the author/owner of the model results desired, and offer co-authorship to the modeler on any publications resulting from the collaboration.

3. Summary and Agenda

To sum up, the tasks confronting the modeling committee, which have been discussed above, can be broadly categorized as involving:
- collection of models, model results;
- disposition of models and results;
- access to models and results.
The steps in development of the database may be summarized as follows:

1. Identify the models
   - categorize
   - contact the owner/developer

2. For each model, decide if the model or model output will be archived
   - for large models, determine which parameters will be archived

3. Use the export format or some other agreed upon format to dispose the models and model output to the database

Construction of the modeling data base can begin with a small subset of the anticipated entries, such as IRI and MSIS and a few TGCM runs. These will constitute a pilot program to develop and test the system.

Issues for committee and community discussion:

- Should CEDAR-funded modelers be required to participate in the CEDAR database?
- What are the rules for use of the model outputs?
- Must contact be made with the model developer?
- Is an offer of co-authorship required?
All words are 16-bit, 2's complement integers.

**Physical Record**

<table>
<thead>
<tr>
<th>Logical record</th>
<th>Logical Record</th>
<th>Logical Record</th>
</tr>
</thead>
</table>

First word: total number of words in this physical record, including this word and checksum.

Each physical record contains an integral number of logical records.

**Logical Record (Data Record)**

<table>
<thead>
<tr>
<th>Prologue</th>
<th>1-D codes</th>
<th>1-D values</th>
<th>2-D codes</th>
<th>2-D values</th>
<th>2-D values</th>
<th>...</th>
</tr>
</thead>
</table>

Or:
- Prologue
- 1-D codes
- 1-D values
- 2-D codes
- 2-D values
- ...
**THE FIRST CEDAR DATA BASE REPORT**

**Prologue**

<table>
<thead>
<tr>
<th>Word Number</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LTOT</td>
<td>Number of 16-bit words in this record, including this one</td>
</tr>
<tr>
<td>2</td>
<td>KREC</td>
<td>Kind of record (1002 for data record in this format)</td>
</tr>
<tr>
<td>3</td>
<td>KINST</td>
<td>Instrument code</td>
</tr>
<tr>
<td>4</td>
<td>KINDAT</td>
<td>Kind-of-data code, pointing to documentation on analysis procedure used</td>
</tr>
<tr>
<td>5</td>
<td>IBYR</td>
<td>Beginning year for data in this record</td>
</tr>
<tr>
<td>6</td>
<td>IBDT</td>
<td>Beginning month/day (MMDD)</td>
</tr>
<tr>
<td>7</td>
<td>IBHM</td>
<td>Beginning hour/min (HHMM)</td>
</tr>
<tr>
<td>8</td>
<td>IBCS</td>
<td>Beginning centisecond</td>
</tr>
<tr>
<td>9</td>
<td>IEYR</td>
<td>Ending year</td>
</tr>
<tr>
<td>10</td>
<td>IEDT</td>
<td>Ending date</td>
</tr>
<tr>
<td>11</td>
<td>IEHM</td>
<td>Ending hour/min</td>
</tr>
<tr>
<td>12</td>
<td>IECS</td>
<td>Ending centisecond</td>
</tr>
<tr>
<td>13</td>
<td>LPROL</td>
<td>Length of this prologue (at least 16)</td>
</tr>
<tr>
<td>14</td>
<td>JPAR</td>
<td>Number of single-valued parameters (0 permissible)</td>
</tr>
<tr>
<td>15</td>
<td>MPAR</td>
<td>Number of multiple-valued parameters (0 permissible)</td>
</tr>
<tr>
<td>16</td>
<td>NROW</td>
<td>Number of entries for each multiple-valued parameters (0 permissible)</td>
</tr>
</tbody>
</table>
**Sample Header Record**

<table>
<thead>
<tr>
<th>Character Portion of the Record Follows:</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLS: 1-8 9-16 17-24 25-64</td>
</tr>
<tr>
<td>CKSRTN C DRWNO IS DATA RECORD WORD NUMBER</td>
</tr>
<tr>
<td>KRECH 3002 HEADER RECORD, VERSION 2</td>
</tr>
<tr>
<td>KINT 20 ARECIBO</td>
</tr>
<tr>
<td>KINDAT 2001 KINDAT DEFINITION FOLLOWS:</td>
</tr>
<tr>
<td>CKINDAT BASIC PARAMETER DATA</td>
</tr>
<tr>
<td>CKINDAT DATA REDUCED USING AO PROGRAM RWD15</td>
</tr>
<tr>
<td>IBYRT 1987 BEGINNING YEAR FOR THIS DATA SET</td>
</tr>
<tr>
<td>IBDT 127 BEGINNING MONTH AND DAY</td>
</tr>
<tr>
<td>IBHMT 1446 BEGINNING UT HOUR AND MINUTE</td>
</tr>
<tr>
<td>IEYRT 1987 ENDING YEAR FOR THIS DATA SET</td>
</tr>
<tr>
<td>IEYRT 1987 ENDING MONTH AND DAY</td>
</tr>
<tr>
<td>IECST 5200 ENDING UT HOUR AND MINUTE</td>
</tr>
<tr>
<td>LPROL 16 DATA RECORD PROLOGUE LENGTH</td>
</tr>
<tr>
<td>JPAR 14 7 NUMBER OF SINGLE-VALUED PARAMETERS</td>
</tr>
<tr>
<td>MPAR 15 15 NUMBER OF MULTIPLE-VALUED PARAMETERS</td>
</tr>
<tr>
<td>NROW 16 15 NUMBER OF ROWS OF MULTIPLE-VALUED PARAMETERS</td>
</tr>
<tr>
<td>C NROW MAY VARY FROM THE ABOVE VALUE IN DATA RECORDS</td>
</tr>
<tr>
<td>KODS(1) 17 60 INTEGRATION TIME FOR THESE DATA 1. S</td>
</tr>
<tr>
<td>KODS(2) 18 115 ALTITUDE AVERAGING INTERVAL 1. KM</td>
</tr>
<tr>
<td>KODS(3) 19 116 ADDITIONAL INCREMENT TO HT AVGNG INTRVL 1.E-01 M</td>
</tr>
<tr>
<td>KODS(4) 20 130 MEAN AZIMUTH ANGLE (O=GEOG N,90=EAST) 1.E-02 DEG</td>
</tr>
<tr>
<td>KODS(5) 21 140 ELEVATION ANGLES (O=HORIZONTAL,90=VERT) 1.E-02 DEG</td>
</tr>
<tr>
<td>KODS(6) 22 535 LOG10 (MAX NE IN M-3) 1.E-03 LOG(M-3)</td>
</tr>
<tr>
<td>KODS(7) 23 540 HEIGHT OF MAXIMUM ELECTRON DENSITY 1. KM</td>
</tr>
<tr>
<td>KODM(1) 31 110 ALTITUDE (HEIGHT) 1. KM</td>
</tr>
<tr>
<td>KODM(2) 32 111 ADDITIONAL INCREMENT TO ALTITUDE 1.E-01 M</td>
</tr>
<tr>
<td>KODM(3) 33 520 LOG10 (NE IN M-3) 1.E-03 LOG(M-3)</td>
</tr>
<tr>
<td>KODM(4) 34 -520 ERROR ESTIMATE FOR PARAMETER 520 1.E-03 LOG(M-3)</td>
</tr>
<tr>
<td>KODM(5) 35 550 ION TEMPERATURE, TI 1. K</td>
</tr>
<tr>
<td>KODM(6) 36 -550 ERROR ESTIMATE FOR PARAMETER 550 1. K</td>
</tr>
<tr>
<td>KODM(7) 37 560 ELECTRON TEMPERATURE, TE 1. K</td>
</tr>
<tr>
<td>KODM(8) 38 -560 ERROR ESTIMATE FOR PARAMETER 560 1. K</td>
</tr>
<tr>
<td>KODM(9) 39 580 LINE OF SIGHT ION VELOCITY (POS = AWAY) 1. M/S</td>
</tr>
<tr>
<td>KODM(10) 40 -580 ERROR ESTIMATE FOR PARAMETER 580 1. M/S</td>
</tr>
<tr>
<td>KODM(11) 41 660 COMPOSITION - [H+] / [O+] 1. E-03</td>
</tr>
<tr>
<td>KODM(12) 42 -660 ERROR ESTIMATE FOR PARAMETER 660 1.E-03</td>
</tr>
<tr>
<td>KODM(13) 43 410 SIGNAL TO NOISE RATIO 1. E-02</td>
</tr>
<tr>
<td>KODM(14) 44 430 GOODNESS OF FIT 1.</td>
</tr>
<tr>
<td>KODM(15) 45 457 ARECIBO DATA QUALITY CODE 2 1.</td>
</tr>
<tr>
<td>C ASSUMED PARAMETERS ERROR VALUES ARE ASSIGNED A VALUE OF -32766</td>
</tr>
<tr>
<td>C QUALITY CODE 430 IS THE RMSERR * 1000, WHERE RMSERR IS DEFINED AS</td>
</tr>
<tr>
<td>C QUALITY CODE 430 IS THE RMSERR * 1000, WHERE RMSERR IS DEFINED AS</td>
</tr>
<tr>
<td>C QUALITY CODE 430 IS THE RMSERR * 1000, WHERE RMSERR IS DEFINED AS</td>
</tr>
<tr>
<td>C QUALITY CODE 430 IS THE RMSERR * 1000, WHERE RMSERR IS DEFINED AS</td>
</tr>
<tr>
<td>C QUALITY CODE 430 IS THE RMSERR * 1000, WHERE RMSERR IS DEFINED AS</td>
</tr>
<tr>
<td>C QUALITY CODE 430 IS THE RMSERR * 1000, WHERE RMSERR IS DEFINED AS</td>
</tr>
<tr>
<td>C QUALITY CODE 430 IS THE RMSERR * 1000, WHERE RMSERR IS DEFINED AS</td>
</tr>
</tbody>
</table>

**CANALYST DATA COLLECTED AND REDUCED BY C. TEPLEY AND R. BURNISDE**

**CEXPER GITCAD (GLOBAL IONOSPHERE-THERMOSPHERE COUPLING AND DYNAMICS)**
### Sample Data Record

#### LOGICAL RECORD NUMBER 31729

**PROLOGUE:** 270 1002 20 2001

**BEGIN**

<table>
<thead>
<tr>
<th>yr</th>
<th>mlday</th>
<th>hr</th>
<th>mn</th>
<th>cs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>129</td>
<td>1628</td>
<td>4800</td>
<td></td>
</tr>
</tbody>
</table>

**PROLOGUE:** 16 7 15 15

#### 1D CODES:

<table>
<thead>
<tr>
<th>km</th>
<th>1m</th>
<th>.01°</th>
<th>.01°</th>
<th>km</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>115</td>
<td>116</td>
<td>130</td>
<td>140</td>
</tr>
</tbody>
</table>

#### 1D VALUES:

<table>
<thead>
<tr>
<th>km</th>
<th>1m</th>
<th>1000 x</th>
<th>Tn</th>
<th>error</th>
<th>Tc</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>644</td>
<td>54</td>
<td>437</td>
<td>31</td>
<td>-1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

#### 2D CODES:

<table>
<thead>
<tr>
<th>km</th>
<th>1m</th>
<th>1000 x</th>
<th>Tn</th>
<th>error</th>
<th>Tc</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>644</td>
<td>54</td>
<td>437</td>
<td>31</td>
<td>-1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

#### 2D VALUES:

<table>
<thead>
<tr>
<th>km</th>
<th>1m</th>
<th>1000 x</th>
<th>Tn</th>
<th>error</th>
<th>Tc</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>644</td>
<td>54</td>
<td>437</td>
<td>31</td>
<td>-1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
PROPOSED CEDAR DATA BASE "RULES OF THE ROAD"

Smooth functioning of the data base requires that there be clear agreements among the parties involved in acquiring, handling, and using the data. The scientists who submit data have invested considerable time, effort, and expertise in collecting and processing the data for submission to the data base. Despite this effort, there are still uncertainties and limitations of the data, making it important for the user to contact the data suppliers early on in a project. The suppliers will help the user understand the characteristics and limitations of the data, and may even be willing to collaborate in prospective studies. It is important that these efforts receive appropriate acknowledgement by users of the data. In addition, the data base needs to maintain records to evaluate how it is being used. The following “Rules of the Road” are proposed to satisfy these needs and to clarify the responsibilities of users.

1. The prospective user must submit a Project Form to obtain access to the data base. Projects forms must be updated at least annually in order to keep a current log in account.

2. Data obtained from the data base are to be shared only with other users who have an up-to-date project form on file with the central facility.

3. The user is required to establish early contact with the organization(s) whose data are involved in the project to discuss the intended usage, in the light of possible data limitations.

4. Before they are formally submitted, draft copies of all reports and publications must be sent to the contact scientist at the data-supplying organization(s) along with an offer of co-authorship to scientists who have provided data. This offer may be declined.

5. The data base and the organizations that contributed data must be acknowledged in all reports and publications.

6. Copies of reports and papers are to be sent to the data base scientist so that the
THE FIRST CEDAR DATA BASE REPORT

project file can be kept up to date.
ASTRONOMICAL DATA ANALYSIS FROM REMOTE SITES

Supriya Chakrabarti, Carl Dobson, George Kaplan, Herman Marshall, Michael Lampton, Roger Malina, Stuart Bowyer

Space Sciences Laboratory, University of California, Berkeley, CA 94720, U.S.A.

Jeff Star

Remote Sensing Research Unit, University of California, Santa Barbara, CA 93106, U.S.A.

Abstract

Given the progress in the communication technology, it is expected that during the space station era the mode of instrument operation and data analysis will be dramatically different from those at present. A consortium of several universities and NASA centers are evaluating various aspects of design and operation of scientific instruments and data analysis over various computer networks from a remote site. Such a scheme has officially been termed tele-science. We will report on the development of methodologies for teledesign, teleoperation and teleanalysis and the verification of these concepts using the Extreme Ultraviolet Explorer (EUV), a satellite payload scheduled for launch in 1991. The EUVE telescopes will be operated remotely from the EUVE Science Operation Center (SOC) located at the University of California, Berkeley. Guest observers will remotely access the EUVE spectrometer database located at the SOC.

Distributed data processing is an integral part of tele-science. We will describe our experience with the Browse system, currently being developed at the University of California at Santa Barbara through a grant from NASA for remote sensing applications. We will discuss the suitability for its adoption for astronomy applications. Browse allows the examination of a subset of the data to determine if the data set merits further investigation. The examination can be as simple as looking for a specific data element based on its location, date of observation, quality indicator, spectral coverage etc. It also allows the viewing of data in various modes depending upon the available resources at the user's end (e.g., graphics terminal vs. dumb terminal), level of data compression applied, required display format etc. and its transmission over a network to a local graphics display station.

Introduction

With the availability of sensitive array detectors the data rates for the modern astronomical instruments have been growing tremendously over the past few years. Even with the availability of high speed computers, the processing of the voluminous data these instruments generate are tackled only on one experiment at a time. During the space station era, we anticipate a large number of collaborative, multi instrument experiments conducted from geographically distributed sites which will easily exceed the database of typical experiments conducted today. Furthermore, since most of these experiments will be conducted by astronomers from different institutions, a new mode of science operation will need to be employed, which will further complicate the data analysis.

It is likely that the database generated by these instruments will be maintained in in several sites. This distribution of the database implies a need for efficient data transfer between the sites for analysis purpose. "Electronic browse" of the available data set will help minimize the required data transfer. Furthermore, it is obvious that each site will have certain unique expertise, analysis tools etc. which need to be shared between the collaborating scientists.

The Office of Science and Application (OSSA) of NASA has recognized these needs and has set up a set of testbed activities to evaluate the feasibility of remote operation of instruments and correlative data analysis from the experimenters' home institution. This new mode of conducting experiments and data analysis has been termed tele-science. An important goal of these activities is the identification of critical issues for the design of the Science Applications Information System (SAIS) and the Space Station Information System (SSIS). We are involved in several of these activities at the University of California at the Berkeley (UCB) and Santa Barbara (UCSB) campuses. The Berkeley efforts use various aspects of an Extreme Ultraviolet (EUV) satellite mission, while the Santa Barbara work is focussed primarily on remote sensing data. In this paper we describe our joint efforts to explore the possibility
of adapting a distributed database system developed at UCSB for use with UCB's EUV satellite data.

The Extreme Ultraviolet Explorer (EUV)

The EUVE instrument consists of three scanning telescopes, which will perform an all-sky survey in the extreme ultraviolet (80-800 Å) and a deep survey/spectrometer telescope that will collect imaging and spectroscopic data along a narrow band near the ecliptic (Bowyer et al., 1986). Raw data will consist of lists of time-tagged photon detections, which are then remapped to sky positions using spacecraft attitude data. At the Space Sciences Laboratory (SSL) of UCB a network of sun workstations will be used for data analysis. This is the same network that is currently in use for development of the analysis software, as well as for instrument design, test, and calibration.

A major goal of the Telescience initiative is for experimenters to be able to operate their (possibly space-based) instruments from their home institutions as well as analyze the data obtained from one or more experiments. With the increasing number of collaborative research efforts involving widely separated sites, it will also be important to develop methodologies for these institutions to cooperate on the design of the experimental instruments, and data analysis schemes. These methodologies, teleoperation and teledesign, are the focus of three Telescience testbed activities involving the EUVE project at UCB.

In the area of teleoperation, an experiment is underway to simulate the operation of the EUVE science instruments from a remote site using one of the existing TCP/IP networks. The instrument are represented in these tests by a "Kiwi" (a "flightless bird") consisting of breadboard electronics and other prototype hardware. It is connected to a single-board computer in essentially the same configuration as will be used with the real instruments during orbital operations. This workstation receives telemetry and sends commands to the onboard computer of the instrument. Software to relay instrument commands over a network using the TCP/IP protocols has been written to allow a second workstation to be used as the command interface. Testing will proceed in steps over progressively longer distances. Initial tests, using two workstations on the SSL ethernet, have already been performed. Subsequent tests will be performed with the second workstation at Stanford University, using the Bay Area Regional Research Network (BARRNET), which connects several research sites in the San Francisco Bay Area using high speed (T1) links.

In the final system the two workstations will be on opposite coasts, with the ground support workstation at GSFC and the SOC workstation at UCB-SSL. The actual link between the SOC and GSFC could be either a dedicated line or a connection through an existing network such as ARPA-net or NSF. A network connection has certain advantages over a dedicated line, but new concerns over command validation arise due to the large number of users who have access to the networks. The goal of the teleoperation experiments is to help answer these concerns as well as others about whether the network can supply the bandwidth necessary for control of the EUVE instruments.

In the teledesign area, we are designing a scheme which allows software development by a Software Control System (SCS), that allows programmers and users to coexist peacefully. That is, versions of many of the analysis software modules are already in use for instrument testing and calibration, at the same time as more advanced versions of the same modules are under development. A "promotion" scheme provides a controlled mechanism for installing revised software modules without causing unpleasant surprises for the users.

In order to encourage the sharing of software resources, all software and associated files are globally accessible from anywhere in the SSL computer system. This is not accomplished with a central computer or file server. Instead, each workstation has its own disk, with files being accessible from any of the other workstations, using San Microsystems' Network File System (NFS). From the perspective of teledesign, it is natural to extend the locally shared software environment to be used between institutions on a wide area network such as the ARPA-net. As part of the Telescience Testbed Program, the SCS is being extended to provide such a service, but replacing the NFS file-sharing facility with a "borrow" mechanism that propagates updated versions of modified files at daily intervals.
The Brow System

As a part of the telescience testified activity an electronic browse capability is being developed at
the Santa Barbara campus of University of California. The Brow System utilizes a database consisting
of images and texts with the goal of allowing the user to preview the database from remote locations.
The primary database used by the Santa Barbara group is remote sensing data from Landsat and
other sources. The key goal of this testbed is to demonstrate the viability of a distributed database sys-
tem using local database and specialized processing software with users accessing it through wide area
computer networks. Preliminary results detailing experience with the current Brow System implementa-
tion are outlined elsewhere (Star et al., 1987a).

The Brow System is aimed at defining methodologies for managing large volumes of data
expected from future missions (estimated at a rate of one day for the Earth Observing System (EOS), for
example, Arvidson et al., 1986). With such large databases, the ability to electronically browse the
database before retrieval is essential. The present estimate of the number of user requiring such browse
capability ranges from 1500 to 15,000 (Star et al., 1987a). It is reasonable to expect that the familiarity
of these users with the database will vary widely. Therefore a need exists for several levels of user
interface to the database.

In order to explore these issues a prototype system consisting of an image and other spatial data
and descriptive information about the data (metadata) has been created which serves as one node of the
distributed image database system. The present database runs on a Digital Equipment Corporation
microVAX II minicomputer running the VMS operating system. The database management software
being used is RIM, a relational database management system (DBMS) created by the Boeing Aerospace
Corporation. An IBM PC/AT microcomputer is used as a remote user workstation with a telephone line
using standard communication protocol as the communication network.

In this experiment the microVAX is being used as a dedicated database computer. A set of utili-
ties has been written for use on the remote user workstation which includes communication software,
graphics software for defining the geographic area of interest as well as image display software which
allows the user to view subset images of selected data sets (Star et al., 1987b). These utilities, in con-
junction with the database software at the host node, allow the user to locate and preview potentially
useful data sets for a specific application. For example, a geographic region can be described using the
interactive graphics software on a map displayed on the user's workstation. After the area of interest is
chosen, available data sets for that region can be selected based upon several constraints such as date,
sensor type, cloud cover and other scene-specific criteria.

As shown in Figure 1, each node of the database consists of three major components, the direct-
tory, the catalog, and the user databases. The directory is a database containing information regarding
the nodes in the distributed database and a description of the contents at each node. It is resident in
each node of the network and contains one entry for each node in the network. The entries describe
the nodes with a set of attributes such as, the number of data objects available, the geographic and scientific
emphasis of the node, and the data format.

The catalog, which is unique to each node, consists of detailed attributes of the data, which can
be used by users to select appropriate images. These attributes include the form of the data such as,
film, printout, digital etc., the description of the data such as, date and time of acquisition, platform,
sensor, spectral coverage etc., and quality indicators such as, cloud cover, sun angle and any processing
history which may have been performed on the data.

The user database contains information regarding local users, such as their field of specialization.
Since the current implementation is based on remote sensing applications, it contains the user's
scientific and geographic specialization. It also contains information regarding system usage, thereby
allowing the users to obtain information regarding other scientists involved in related research.

In order to evaluate the performance of this strawman system a group of scientists has been
selected to use it. Since no network links are currently in place, a toll free telephone number has been
established to allow the test users to access the system at UCSC. Through this link the users can query
the database, download the required image data and then perform graphic manipulations at their local
workstations.
A typical session with the present system will involve specifying a set of attributes through a menu-driven procedure. The geographic area of interest is next identified interactively by drawing a window on a map with a pointing device. The user is then prompted for more specific information. For example, a climatologist interested in the land surface energy balance research in the Santa Barbara area might enter the following attributes:

**Database of interest:** CATALOG
**Platform:** ALL
**Sensor:** ALL
**Begin Date:** June 1, 1979
**End Date:** Sept 1, 1979

so that the system might respond with information of the following nature:

**Browse Database:** 2 items found

1. Heat Capacity Mapping Mission,
   July 3, 1979
   3 bands on line, 600m spatial resolution
   scene id # A-A0062-16032-2

2. Landsat-3,
   August 15, 1979
   4 bands on line, 600m spatial resolution
   scene id # 8246517502500

For more information please enter the scene id or ESC to escape:

At this point the user may obtain more detailed information such as the sensor specifications, resolution, number of lines, sun angle, etc. by entering the scene id. This is how the user "browses" through the image database and identify potentially useful data sets. When an appropriate image is found, the user exits the database to download the image data at the remote workstation. The image display software can then be used to manipulate the actual image data.

**The Adaptation of Browse to EUVE**

One of the important components of the EUVE program is the Guest Observer (GO) program for spectroscopic observations with the Deep Survey/Spectrometer Instrument. The current plans for the GO data analysis include a local database and distributed scientific users having various levels of access to the database. These components are all similar to those in the Browse system. Furthermore, the EUVE data makes heavy use of images of astronomical objects, another similarity with the Browse system. Finally, the spectroscopic data are obtained by imaging spectrometers, and so they can be treated like images. For all these reasons we are exploring the possibility of incorporating the Browse system or some of its key components in the EUVE GO data analysis system. The following section describes the GO data handling plans and its similarities with the Browse scheme. Since the EUVE GO plans are in their infancy and the Browse is not available for external users, we describe only our current plans for the GO data analysis.

**The EUVE Guest Observer Data Analysis Plan**

Data processing for GO's will proceed in much the same manner as with sky survey data (see Marshall et al., these proceedings). At the discretion of the guest observer, the off-line mode of operations will require permanent files that are generated in the production processing. Binned photon and exposure maps are generated that may be deconvolved for some studies (e.g. extended sources and crowded fields), while detailed photon and exposure data are saved in "pigeon hole" files for studies of individual sources. It is expected that the latter data set will be more commonly used. The collection of detailed data into pigeon holes (as applied to spectra) is governed by the list of sources detected in
the deep survey instrument which provides a direct image of the target region. These data will be similar in format to those in the Browse database. As with the survey data, sources may be found which were not expected or which may be brighter than expected and thereby contribute to noise in the spectrum. Thus, off-line analyses are required to collect data for sources that were not in the original catalog. This off-line processing should run in a fashion similar to that of the scanner data (Marshall et al., Figure 4). Data for newly discovered sources will be collected from prior observations for future analysis.

The main differences between the handling of survey and spectroscopic data are in the treatment after collection into a permanent database. Figure 2 shows a preliminary scheme for handling GO data. Generally, all data are handled by the GO Interface software which allows data and software transfers via phone lines, networks, or physical media (e.g., optical disks for large requests or printout for users without access to computer networks). Browse has schemes that account for the different levels of access to the database; even different types of user environments e.g., dumb terminal vs. graphics workstation interfaces are also included in their interface. We will examine these schemes before selecting the GO interface.

The GO Interface selects the applicable portion of the database to minimize file transfer time and should warn the remote user if the requested data are not in the current database. In such a case, a special request is generated that requires action by the database manager, which controls the archive data. Requests for unavailable data are handled by generating a set of new coordinates for guiding the selector program via the database manager. This process assures that the raw data are managed properly by an on-site operator of the database managing program. Unlike Browse, we plan to allow remote logins but access to the production network is restricted for security. Temporary files may be created by running remote processes and then transferred to the GO’s home institution.

The GO database plan calls for a file called history. It contains general orientation information, instrument housekeeping information and other parameters such as sun angle, positions of the moon and other planets, etc. There are also plans for storing intermediate results such as exposure information, photon maps, data quality indicators, etc. This database is therefore essentially the same as catalog in the Browse system, which contains information that describes the quality of the data and other attributes.

It is expected that a number of specialized I/O and data analysis packages will be written specifically for use in the IRAF data analysis environment, an image analysis package developed by the National Optical Astronomy Observatories (NOAO). These packages will be available for remote execution or transfer to the GO's home institution for installation in IRAF. This is similar in concept to the software Browse will provide to its remote users for manipulation of images. The packages supplied for EUVE data analysis will run on any hardware configuration to which IRAF has been ported. Using IRAF's potential networking functions, one may also invoke this package at the remote site without transferring the applicable files.

We also anticipate several simultaneous observations with EUVE by other instruments (ground- or space-based). These GO nodes will contain special data not in the EUVE database as well as specific scientific interest and expertise. These characteristics are also incorporated in Browse and are included in the directory and user databases. We anticipate that the EUVE GO database management plan will include similar features.

Unfortunately Browse, as it is presently implemented, uses a VMS® operating system. All of the EUVE software is built around a network of SUN workstations running Unix®. Our SCS scheme, described earlier makes heavy use of UNIX facilities. Therefore, we will not be able to incorporate Browse readily. We will, however, include several key concepts of Browse.

Acknowledgements

We thank Mark Friedl for helpful discussions. This work has been performed under OSFC contract NAS5-29298, NASA subcontract NASW-4234, and the University Space Research Association subcontract 800-58 to the University of California, Berkeley.
References


Figure Captions

1. A schematic of the Browse distributed database management system.

2. A schematic of the EUVE guest observer data management plan.
THE FIRST CEDAR DATA BASE REPORT

Attributes of the node (same format for all nodes)
- Number of available data objects
- Scientific interest of the node
- Principal collections and their format

Detailed inventory of the data holdings
- Image Attributes
  - Date & time of acquisition
  - Process history
  - Platform
  - Sensor
  - Cloud Coverage

Information regarding users
- Geographic Specialization
  - Scientific Specialization
  - System usage information