Data Systems for Science Integration within the Atmospheric Radiation Measurement Program

Deborah K. Gracio
Larry D. Hatfield
Kenneth R. Yates
Jimmy W. Voyles
Pacific Northwest Laboratory, Richland, Washington

Joyce L. Tichler
Brookhaven National Laboratory, Upton, New York

Richard T. Cederwall
Mark J. Laufersweiler
Martin J. Leach
Lawrence Livermore National Laboratory, Livermore, California

Paul Singley
Oak Ridge National Laboratory, Oak Ridge, Tennessee

1. Introduction

The Atmospheric Radiation Measurement (ARM) Program was developed by the U.S. Department of Energy to support the goals and mission of the U.S. Global Change Research Program. The purpose of the ARM program is to improve the predictive capabilities of General Circulation Models (GCMs) in their treatment of clouds and radiative transfer effects. Three experimental testbeds were designed for the deployment of instruments to collect atmospheric data used to drive the GCMs. Each site, known as a Cloud and Radiation Testbed (CART), consists of a highly available, redundant data system for the collection of data from a variety of instrumentation. The first CART site was deployed in April 1992 in the Southern Great Plains (SGP), Lamont, Oklahoma, with the other two sites to follow in early 1996 in the Tropical Western Pacific (TWP) and in 1997 on the North Slope of Alaska (NSA).

Approximately 1.5 GB of data are transferred per day via the Internet from the CART sites, and external data sources to the ARM Experiment Center (EC) at Pacific Northwest Laboratory in Richland, Washington. The Experiment Center is central to the ARM data path and provides for the collection, processing, analysis and delivery of ARM data. Data from the CART sites from a variety of instrumentation, observational systems and from external data sources are transferred to the Experiment Center. The EC processes these data streams on a continuous basis to provide derived data products to the ARM Science Team in near real-time while maintaining a three-month running archive of data.

A primary requirement of the ARM Program is to preserve and protect all data produced or acquired. This function is performed at Oak Ridge National Laboratory where leading edge technology is employed for the long-term storage of ARM data. The ARM Archive

This research was performed under the auspices of the United States Department of Energy under Contract No. DE-AC02-76CH00016.

provides access to data for participants outside of the ARM Program.

The ARM Program involves a collaborative effort by teams from various DOE National Laboratories, providing multi-disciplinary areas of expertise. This paper will discuss the collaborative methods in which the ARM teams translate the scientific goals of the Program into data products. By combining atmospheric scientists, systems engineers, and software engineers, the ARM Program has successfully designed and developed an environment where advances in understanding the parameterizations of GCMs can be made.

2. History

Planning for the ARM Program began in the fall of 1989; a description of the initial program is available in the ARM Program Plan (U.S. Department of Energy 1990). The technical approach of the ARM Program and the design of the CART sites is discussed in more detail by Stokes and Schwartz (1994).

The design of the CART data systems was part of the initial program plan (Melton, et al. 1991). The plan called for a distributed environment, the CART Data Environment (CDE), which included an Archive, an Experiment Center, and a collection of data systems distributed across all of the CART facilities for the collection and processing of data. The CDE was to be implemented "based on use of existing solutions wherever possible, evolutionary implementation of functionality, and parallel implementation of independent subsystems" (Melton et al. 1992). Figure 1 shows the flow of data through the CART Data Environment.

The Southern Great Plains (SGP) CART site was the first to become operational, with instrument deployment beginning in April 1992. Figure 2 is a map of the SGP site and the instruments deployed there.

The site is slightly larger than a GCM grid cell, and is approximately 350 km (N-S) by 250 km (E-W). The central facility contains

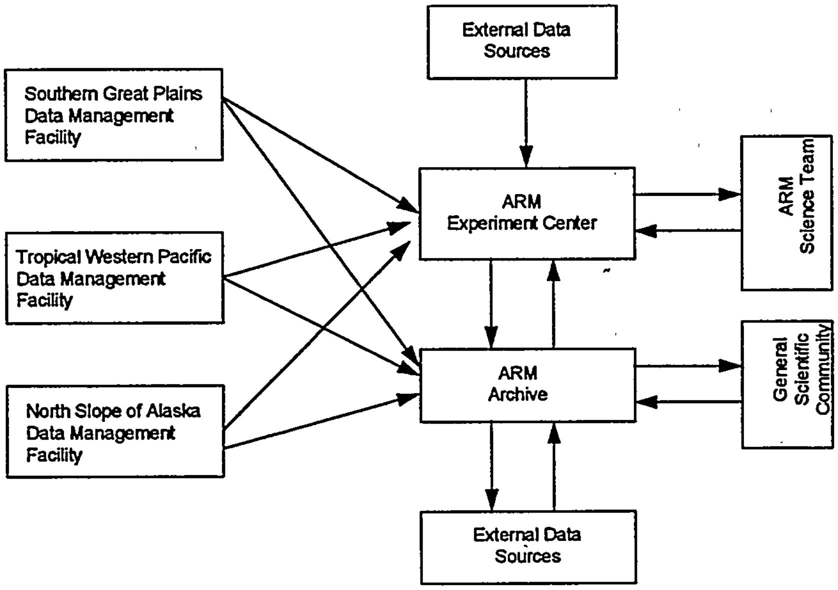
External Data Sources
Southern Great Plains Data Management Facility
Tropical Western Pacific Data Management Facility
North Slope of Alaska Data Management Facility
ARM Experiment Center
ARM Science Team
ARM Archive
General Scientific Community
External Data Sources

Figure 1: Data flow through the CART Data Environment

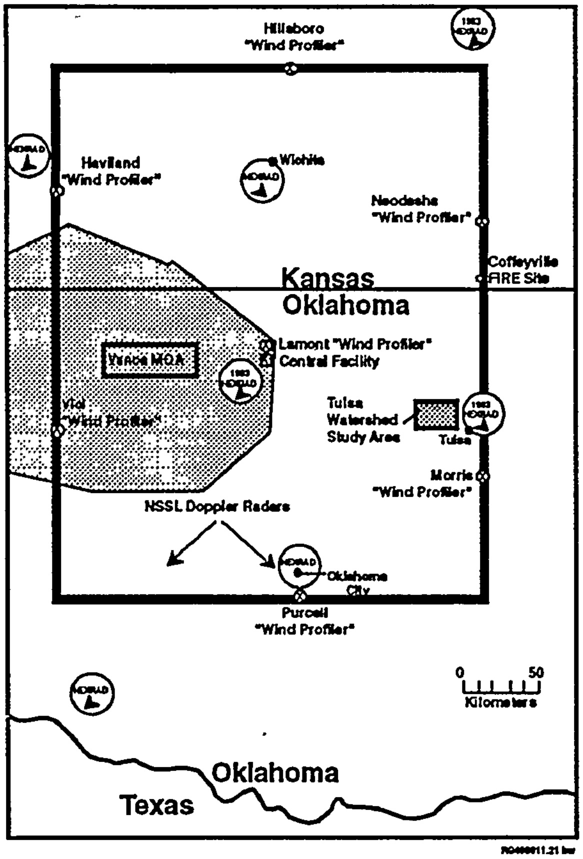

Figure 2: SGP CART site with Instrument Locations

the greatest number and variety of instruments. The principle objective of these experimental testbeds is to quantitatively describe the spectral radiative energy balance profile within a wide range of meteorological conditions. Measurements are taken at a finer spatial resolution than GCMs utilize. The purpose of these measurements is to improve the parameterizations of sub-grid scale processes for use in GCMs. The sites will be operated for 7-10 years, continuously collecting measurements of atmospheric radiation and associated atmospheric and surface properties (Stokes and Schwartz 1994).

3. How the Technical Infrastructure Facilitates the Scientific Activities of the ARM Science Team

Interaction between the Data and Science Integration Team (DSIT) and the ARM Science Team (ST) is a primary mechanism that guides data collection and management within the ARM Program. This interaction leads to the translation of scientific goals and objectives into data requirements that define experiments for individual ST members. Figure 3 is a conceptual diagram of the science-related functions of the ARM technical infrastructure. The flow illustrated in the figure begins with a scientific dialogue between the ARM scientists, collectively or individually, and the ARM infrastructure to define specifically what data are needed to accomplish the stated scientific objectives. From this point, the infrastructure implements the actions necessary to provide the required data to the scientists. Scientific feedback from users of the data is a key feature for improving the quality and usefulness of the data.

ARM is unique in its approach of interacting proactively with funded scientists. The interactive process between the DSIT and the ST is facilitated by a liaison from the DSIT, who understands the ST member's planned research and assists in the identification and acquisition of required ARM data. This process identifies data requirements not currently met by data within ARM. There are several actions that the ARM Program can take to obtain the required data. The acquisition of required data can present difficulties that are not encountered in the management of routine data. In some cases, the required data may be available from a source outside ARM.

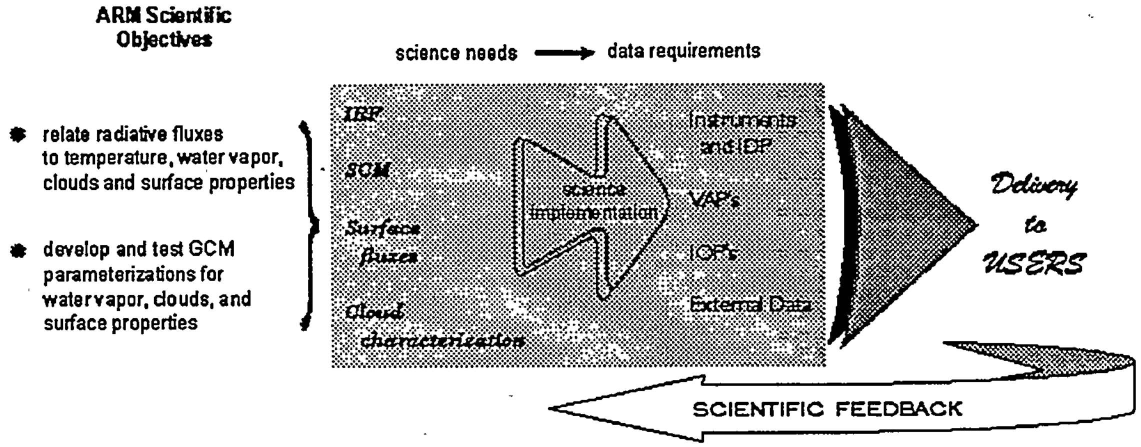
ARM Scientific Objectives
relate radiative fluxes to temperature, water vapor, clouds and surface properties
develop and test GCM parameterizations for water vapor, clouds, and surface properties
science needs
data requirements
IRF
SCM
Surface fluxes
Cloud characterization
science implementation
Instruments and IOP
VAPs
IOPs
External Data
Delivery to USERS
SCIENTIFIC FEEDBACK

Figure 3: Scientific Functions of the ARM Infrastructure

Several steps are taken to enhance the observational capability of the applicable ARM Cloud and Radiation Testbed sites when data requirements point to the need for more observations. This enhancement can sometimes be as simple as the implementation of an Intensive Observation Period (IOP), where the temporal resolution of the observations is increased (e.g., more frequent radiosonde launches). Enhancement may be gained through the acquisition and deployment of additional instrumentation at existing facilities. Finally, the development and testing of new instruments may be required, and this is done within the Instrument Development Program, a component of ARM.

In many cases, data requirements can be satisfied by new calculated data streams. A calculated data stream may be the merger of two or more observations to produce an unobserved quantity (e.g., vertical velocity from the horizontal wind divergence). Algorithms used in calculations may be defined by scientists within or outside ARM or developed within the infrastructure.

One of the strengths of the Science Team concept is the potential for synergism through interaction and information exchange among ST members. This interaction happens most naturally around science issues and related data sets of common interest. To date, the two most active interaction areas have been clear-sky instantaneous radiative flux (IRF) and single-column modeling (SCM). The DSIT is a catalyst for developing such interactive working groups and communicating the needs of the working group to the ARM infrastructure. The identification of "showcase" data sets of common interest helps focus ST and infrastructure attention that increases the benefit of the collected data sets and stimulates progress not possible by ST members working in isolation.

4. ARM Data Sources and Products

ARM data are normally collected in an ongoing, continuous manner, punctuated by Intensive Operations Periods. These two methods of collecting data complement each other. The "first-generation" of a data stream is the observations taken directly from the instrumentation in the field. The quality of the data is assessed and documented via simple minimum, maximum and delta threshold checks. The ongoing nature of regular ARM operations requires an automatic approach in analyzing the data. To this end, the concept of Derived Products has been defined.

A Derived Product is the definition of a procedure which creates a "second-generation" data stream by using existing ARM data streams as inputs and applying algorithms or models to them. The procedure is run automatically and continuously as long as there is input data, and the output (called a "derived product") becomes a new data stream.

Prospective derived products may be identified by any part of the program, from instrument developers to science team members or any place in between. There are two distinct types of derived products. The first type consists of data processing, including smoothing, interpolation, extrapolation, time synchronization of different data streams, and/or time averaging. Another example of this type of product would be a procedure that applies new calibrations to existing data, thus creating a new data stream. These products are designed to either reformat the data to make it easier for ST members or models to use, or to reprocess the original dataset to improve the quality of the data.

The second type of derived product generates new data streams derived either from physical models driven by inputs from existing ARM data streams, or from data quality comparisons. These algorithms come from ST members who are interested in the derived physical quantities or the models.

Quality Measurement Experiments (QMEs) are a subset of the derived products designed specifically to enhance the ARM data quality by providing ongoing, continuous data streams derived from the intercomparison of different ARM data streams. QMEs are part of a two-pronged effort to ensure that ARM data are of known and reasonable quality. The first approach focuses on self-consistency within a single data stream, using various automated methods at the time of processing. QMEs, in contrast compare multiple data streams that are somehow related against a set of expectations. A QME provides the capability to identify data anomalies, such as inconsistent data across instruments, incorrectly implemented or inconsistent algorithms, and the information needed to identify the root cause of these anomalies.

To meet the scientific objectives of ARM, it is necessary to augment observations from the ARM sites with data from other sources. External data sets which are currently being acquired to augment measurements at the Southern Great Plains site are listed in Table 1.

Table 1. External Data Sources

Surface Data
- Oklahoma Mesonet
- Kansas Mesonet
- National Weather Service
- Arkansas Basin River Forecast Center (Stage III Precipitation estimates)

Upper Air Data
- National Weather Service
- NOAA Wind Profiler Demonstration Network

Satellite Data
- GOES
- AVHRR
- TOVS

Numerical Products
- NMC (ETA Model and Rapid Update Cycle RUC)

Some data sets are acquired in near real-time, while other acquisitions are delayed until the responsible organization has had the time to perform quality assurance on the data sets.

Products derived from some of these external data sets or from combinations of the external and ARM data are either currently being generated or are planned. Examples include the cloud coverage statistics derived by a group at NASA Langley from the GOES satellite, and integrated surface mesonet data sets which incorporate the various surface meteorological data stations.

It is anticipated that the need for external data will increase with the addition of the CART sites in the Tropical Western Pacific (TWP) and the North Slope of Alaska. Arrangements are already underway to acquire GMS and AVHRR data for the TWP CART site.

5. ARM Data Systems

The CART Data Environment is made up of three primary components: a CART data system for each experimental testbed, the Experiment Center and the Archive. Each CART site includes a central data system for the acquisition and processing of data. The data systems are designed to run data collection processing of data and data transmission software in a continuous, real-time production environment. Each data system utilizes a software package titled Zebra for data processing activities. Zebra was developed by the National Center for Atmospheric Research (NCAR) to aid in the collection and processing of data during field campaigns (Corbet and Mueller, 1991). All data is stored in a standard data format - NetCDF or Network Common Data Format, developed by Unidata. NetCDF is a binary compressed self-describing data format. The following sections describe in detail the CART data systems.

5.1 Atmospheric Radiation Cloud Stations

To meet the scientific needs of atmospheric scientists, the ARM Program has been faced with the challenge of deploying instruments and sensor suites in remote, and often inhospitable locales. To meet this challenge, the concept of Atmospheric Radiation Cloud Stations (ARCS) was developed. This concept provides for semi-autonomous operation and can be used in a stationary or mobile configuration as necessary. A set of baseline capabilities for the facilities, instruments and data management sub-systems were established. This baseline serves as the acceptance criteria of all elements of the ARCS and provides the design team with a means of measuring reliability and overall quality. The ARCS concept will be deployed at the TWP and NSA sites. A system with very similar requirements was developed for the SGP site early in the program. This system took advantage of the accessibility of Internet and the ability to access the data system locally. The ARCS concept represents the next generation of the CART data systems for the ARM Program.

The overall computing environment for the ARCS is comprised of three primary subsystems. A Monitoring and Control System (MACS), a satellite Communication System (COMMS), and ARCS Data Management (ADaM) system. These three systems are inter-connected via a local area network providing low power, and reliable communication between the systems. An extensible, packet oriented communication protocol has been designed to standardize communication between the systems.

To facilitate command and control of the ARCS systems, a modular external communication system was designed and developed. This system can be used with a variety of satellite and ground based communication mediums. For two-way communication the COMMS system utilizes an Inmarsat-C satellite transceiver. This means of communication will be used to issue commands and queries to the ARCS systems. For health and status reporting the COMMS system uses a GOES transmitter. These GOES transmissions require that data be transmitted on a fixed schedule, therefore advanced preparation of the data to be transmitted is necessary.

Reliable monitoring and control of the systems is crucial to the successful operation of the ARCS. The MACS is designed to be an extremely reliable network of sensors capable of reporting on the status of all components of the ARCS and taking appropriate action to maintain a proper operating environment. The MACS has also been designed to be capable of controlling other sub-systems of the ARCS. In addition, the MACS system can collect instrument data in a degraded mode of operation should the primary data management system fail.

MACS has responsibility for the collection and forwarding of the ARCS health and status reports. Hourly reports of health of each of the monitored aspects of the ARCS are generated by MACS and forwarded to the COMMS system for transmission via satellite. In addition, the MACS requests data from the ADaM system which provides information about the status of the instruments and the quality of the data streams.

The data management system for the ARM TWP Site is faced with a unique set of requirements. In addition to the collection, processing and quality checking of the data collected, the system must also be designed to run with a minimum of human interaction. To accomplish this goal the ADaM System was designed around a "High Availability" architecture (HA). The key to this approach is not to eliminate the possibility of problems but, rather, provide mechanisms to identify and resolve the problems as they occur. If the problems can not be immediately resolved a backup system is always ready to take over. This "self-healing" approach is fundamental to the ADaM design. Other alternative architectures, such as fault tolerant and distributed, were considered however the HA solution was determined to be the most mature and manageable.

The data in the system follow a "critical data path". This path provides for actions to be taken on the data at several locations in the path. As each data set enters a data hold, processes have the opportunity to act on the data before the data set is advanced to the next data hold. The first point that this occurs is in the Data Acquisition and Control Process (DACP) data hold. This is the area where data are received into the system from the instrumentation and are checked for proper format, size and delivery frequency. Next the data are moved into the Incoming data hold, this area is where the raw data are archived onto magnetic media for safe keeping. Once the data set has been archived, it is moved to the Raw data hold. At this point in the path, the data are made available to software modules written to convert data from instrument specific format to the ARM standard netCDF format. After the data have been processed, it is placed in the Processed data hold where it awaits transfer to magnetic tape and if possible, electronic transfer to other ARM data management sites.

As data tapes are received from the ARCS site by the ARCS data management facility, the data is extracted from tape and a series of data quality checks are performed on the data by a site science team. Specific calibration values are applied to the data as necessary and the new data sets are transferred to the ARM Experiment Center and to the ARM archive.

To date, there are 17 different types of instruments and a total of 60 instruments supported by the CART data system installed at the SGP site. The data system is projected to support 22 different types of instruments and a total of 100 instruments by the end of 1996. A subset of these instruments will be deployed with the ARC stations.

5.2 Experiment Center

The ARM Experiment Center (EC) provides a production computing environment for the collection, analysis and delivery of atmospheric data to ARM Science Team members around the country. Data is received into the EC on an hourly basis from the CART facilities and external data sources. Approximately 1.5 GB of data is transmitted daily via the Internet and an estimated 2 GB of data will be received via magnetic media from the TWP and NSA sites when they are deployed. Software applications are run in the EC to determine the quality of the data and to create the input data streams for the climate models. Customized data products are delivered via the Internet to ARM Science Team members based upon their specific needs and experiment requirements.

The Experiment Center consists of a series of multi-processor Sun Sparc stations connected to a 300 GB RAID system. An Erasable Optical Disk storage system is used for maintaining a three month archive of data locally. High speed computers, such as a HP 735 are used for running software applications which provide in depth quality analysis of data, perform data fusion of surface observations, and assimilate vertical profiles of meteorological variables into three dimensional grids. High availability systems are being considered for deployment in the EC to maximize the ability to deliver data streams to the ARM Science Team in an uninterrupted, timely fashion.

5.3 Archive

The ARM Archive stores, manages, and makes data collected during the ARM Program available for use by the ARM Science Team members and the general scientific community. The mass storage system is a collection of robotically controlled tape systems and RAID disks controlled via the National Storage Laboratory (NSL) - Unitree storage management software. NSL-Unitree manages the migration of files through the collection of storage devices transparent to the rest of the archive system. The current capacity of the Archive is approximately 10 TB with growth capability to 300 TB within the next two years.

A graphical user interface is provided so that users may directly request data based on the set of instruments that collect data and the location of the collection. The user interface is supported by a Sybase Relational Database Management System that contains the metadata for each data file and a pointer into the storage device for location of the data file.This information is primarily designed to assist the users in understanding what data sets are available. The requested data files are made available to users for pick-up via ftp.

6. Challenges

The ARM DSIT team is made up of a diverse set of disciplines including electrical, mechanical, and software engineers, mathematicians, physicists, and atmospheric scientists. Members of the team are located at five different national laboratories, thus, increasing the requirements for enhanced technological communications. Electronic mail, phone and video conferencing, and travel have become significant contributors to communication among team members.

The group responsible for designing and implementing the ARM SGP data system faced a number of challenges. Specifically, the implementation of a complex distributed system in a relatively short amount of time, with team members located in three separate time zones. Furthermore, the SGP site was not close to any of the laboratories at which the team worked and since it was located in the middle of farm land, the site was started with no communication facilities. The team was designing a system which was to run continuously for an estimated 10 years and be capable of near real-time delivery of data to its ultimate users, the ARM Science Team.

The ARM Program Management developed a general method of tracking and resolving problems by means of a subgroup, the Problem Review Board (PRB), which meets via weekly conference calls. A database was developed to assist the PRB by recording all reported problems via Problem Identification Forms (PIFs) and the resolution of problems via Corrective Action Reports (CARs). In addition, the quality of data streams are documented and stored in the database in the form of Data Quality Reports (DQRs).

7. Collaboration with Researchers and Research Programs outside ARM

Collaboration occurs on both the individual and program level. In many instances, scientists outside ARM conduct research related to ARM scientific objectives. The exchange of data and scientific ideas is mutually beneficial to both parties. In some cases, ARM has strengthened these collaborations by identifying the scientists as adjunct ST members. In other cases, ARM has established funding arrangements for providing desired data to ARM (e.g., NASA providing satellite-derived cloud products).

Collaboration with ongoing research programs is an important part of the ARM approach for meeting its scientific objectives. Since ARM addresses a specific part of the overall global climate problem, other programs can provide scientific understanding, observational approaches, algorithms, and data that enhance the results of ARM-funded research. Also, by combining ARM resources with those of other programs, certain scientific goals can be achieved that individual programs could not achieve on their own (e.g., ARM and SHEBA (Surface Heat Budget of the Arctic Ocean) interactions in the Arctic). The density of instrumentation and the long-term period of data collection at ARM CART sites have attracted several programs that wish to take advantage of the ARM sites as benchmarks. Examples of collaborations include the Global Energy and Water Experiment (GEWEX) through its GCIP subprogram in the Southern Great Plains, and the Tropical Ocean Global Atmospheric - Coupled Ocean Atmospheric Regional Experiment (TOGA-COARE) program in the Tropical Western Pacific.

8. Future Direction of the DSIT

Translating science needs into data requirements and delivering data streams of known and reasonable quality are fundamental principles of the ARM program. Maturity of the capability to realize these principles will enhance the scientific productivity of the ARM Science Team.

The DSIT must keep pace with the increasing capacity of ARM Science Team members to make progress towards the ARM programmatic objectives to improve General Circulation Models and understand the atmospheric radiation-cloud radiative feedback. To accomplish this, the DSIT will maintain a vigorous policy of upgrading the software and hardware data systems and optimize the critical loop between applied science and modeling efforts with those of the ARM Science Team. In essence, the goal is to make high quality data, information, and derived products available to ARM scientists.

Focus groups within the DSIT are working to develop effective methods for designing derived products, data system development activities, and the operation and maintenance of the data environment. These focus groups, along with the infusion of new technologies, and the utilization and re-use of previously developed tools, are moving us toward an open-architecture approach to product delivery, processing, tracking, and characterization of data streams. In particular, the use of the World Wide Web for cross platform access to data and information, object-oriented database techniques to manage meta-data relations at the archive, an integrated development, operations, and maintenance approach, and standard data analysis display tools will continue to have a positive impact.

In the future, as we develop new system requirements and plan the integration of new technologies and algorithms, we will keep the principles and scientific objectives of the ARM Program in view. Understanding the needs of ARM scientists and how well we meet those needs form the basis of the operational measures of the effectiveness of the DSIT.

References

Corbet, J. M., and Mueller, C. 1991. *Zeb: Software for data integration, display and analysis*. Proceedings of the 25th Int. AMS Conference on Radar Meteorology, Paris, France, American Meteorological Society.

Melton, R. B., Campbell, A. P., Edwards, D. M., Kanciruk, P., Tichler, J.L. 1991. *Design of the CART Data System for the U.S. Department of Energy's ARM Program*. American Meteorological Society Proceedings.

Stokes, G. M., and Schwartz, S. E. 1994. *The Atmospheric Radiation Measurement (ARM) Program: Programmatic Background and Design of the Cloud and Radiation Testbed.* Bulletin of the American Meteorological Society. Vol 75, No. 7, 1201-1221.

U.S. Department of Energy. 1990. *Atmospheric Radiation Measurement Program Plan.* DOE/ER-0442.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.