Chapter 11

The ARM Data System and Archive

RAYMOND MCCORD

Oak Ridge National Laboratory, Oak Ridge, Tennessee

JIMMY VOYLES

Pacific Northwest National Laboratory, Richland, Washington

1. Introduction

Every observationally based research program needs a way to collect data from instruments, convert the data from its raw format into a more usable format, apply quality control, process it into higher-order data products, store the data, and make the data available to its scientific community. This data flow is illustrated pictorially in Fig. 11-1. These are the basic requirements of any scientific data system, and ARM's data system would have to address these requirements and more. This chapter provides one view of the development of the ARM data system, which includes the ARM Data Archive, and some of the notable decisions that were made along the way.

It is impossible to talk about the development of the ARM data system without first placing it in context of the evolution of computers and associated infrastructure. At the start of ARM in 1990, the Intel 486 computer with 25-MHz processing speed had just been released, internal hard drives were about 40–100 MB in size, national networks were very loosely connected but had significantly less reliability and capability than today's Internet, and the World Wide Web technology was an experimental concept. However, it was already clear that computers and associated technology were developing at a rapid pace and that any design of the ARM data system would have to be flexible enough to accommodate new technology as it came along. Balancing the need for flexibility was the assumption that ARM was initially envisioned to be a 10-yr research program (Stokes 2016, chapter 2; Cress and Sisterson 2016, chapter 5), and thus at the beginning it was anticipated that the data system might only go through one generation of updates. The growth of the program into a multidecadal program with significantly increasing computational and storage demands placed tremendous stress upon the ARM data system, causing it to be reorganized and reconfigured several times. Today, the data system continues to be a living system, evolving as needed to support the ARM mission and user community.

Figure 11-1 shows a phased and linear view of the ARM data flow. However, the data flow is related to the data life cycle, which occurs when researchers make discoveries with data and identify new science questions and data needs. These questions often require new measurements, new sampling schemes, new advanced data products, and new evaluation procedures (see Fig. 11-2). Because of the close coupling of the science and infrastructure in the ARM Program, the data system was continually adapted to support the new data requirements. This cyclic pattern of change is common in scientific data systems but is often insignificant when the project duration is only 3–5 yr. The ARM data system is unusual since it has undergone a 20-yr history of continual evolution to support the significant increases in data volume, types of data, and configurations of field sites. In addition, the interactions between the researchers and the data system changed in ways that could not be anticipated at the start of ARM.

2. Initial requirements

Data system–related activities for ARM were a major component of the programmatic scope from the very

Corresponding author address: Raymond McCord, Oak Ridge National Laboratory, 1 Bethel Valley Rd., Oak Ridge, TN 37831.
E-mail: mccordra@ornl.gov

DOI: 10.1175/AMSMONOGRAPHS-D-15-0043.1

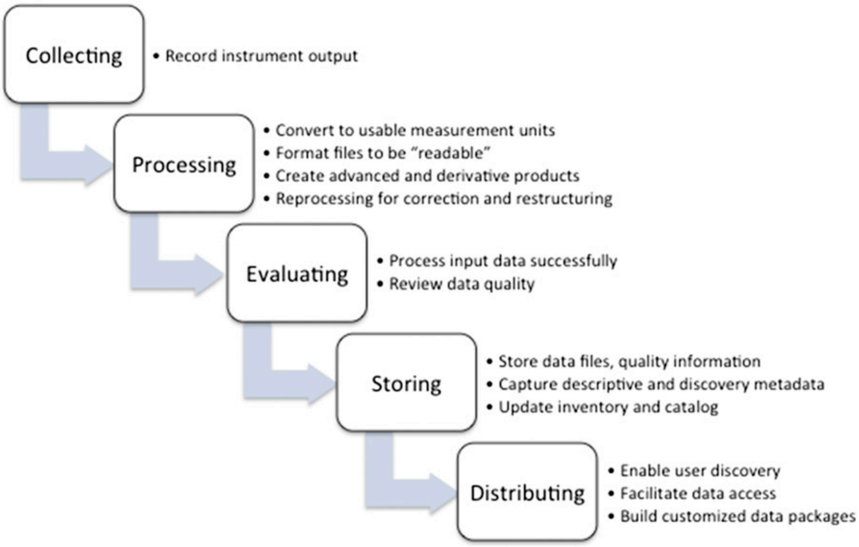
Collecting
• Record instrument output
Processing
• Convert to usable measurement units
• Format files to be "readable"
• Create advanced and derivative products
• Reprocessing for correction and restructuring
Evaluating
• Process input data successfully
• Review data quality
Storing
• Store data files, quality information
• Capture descriptive and discovery metadata
• Update inventory and catalog
Distributing
• Enable user discovery
• Facilitate data access
• Build customized data packages

FIG. 11-1. Functional scope of the ARM data system.

beginning of the planning and implementation (Cress and Sisterson 2016, chapter 5). The major components of the original ARM Program plan (U.S. Department of Energy 1990; ARM 2016, appendix A) included

- the ARM science team (researchers);
- instrument team (acquisition, implementation, and development);
- site operations (development and operations); and
- data systems (design, development, and operations).

In this discussion, data system refers to all portions of the hardware, software, data processing, and data flow. The ARM Program's longstanding research objective of comparing atmospheric process models with measurements for improving models was critically dependent on the data system. Early inclusion of the data system in the programmatic planning process enabled the ARM to define durable requirements and goals (Cress and Sisterson 2016, chapter 5). This very early planning strategy assured readiness of the systems when the first site was operational.

The early period of design and implementation of the data system spanned 1990–93. The early requirements for the ARM data system were divided into the following:

- data life cycle: the flow and processing of data (Table 11-1);
- data heterogeneity: structure, size, and source (Table 11-2); and
- design strategy: arrangement of hardware and software, and development approach (Table 11-3).

The overall structure of the data system was originally divided into three logical groups of systems: site data systems, a data processing facility, and the ARM Data Archive. The data systems remain in these logical groups after the 20 yr of ARM history (Fig. 11-3). The details of the data systems evolved with the maturity of the experimental design and the reality of the first budgets, early instruments, and first site operations.

3. Early advantages and unknowns

The maturity of scientific data management was an important influence on the initial and early designs of the ARM data system. Many of the original ARM data system staff had 5–15 yr of experience from smaller and shorter research programs. They had a good vision for the overall scope of the data system and a good sense of feasibility for the options. The early data managers also knew that an incremental design approach was needed for successful implementation.

At its inception, the size of the ARM data system was near the upper limits of large systems for environmental research. The primary unknowns in the initial data system design were as follows:

- How do we organize, store, and access this very large data volume?
  - Online data storage was very efficient (high speed, easy access) but very expensive with short technology life cycles.
  - Offline data storage was cost effective, but data access was slow and labor intensive. Automated tape libraries were new and had limited reliability.
- How do we transfer the very large data volumes between the field sites, processing centers, quality reviewers, and researchers in a timely manner?

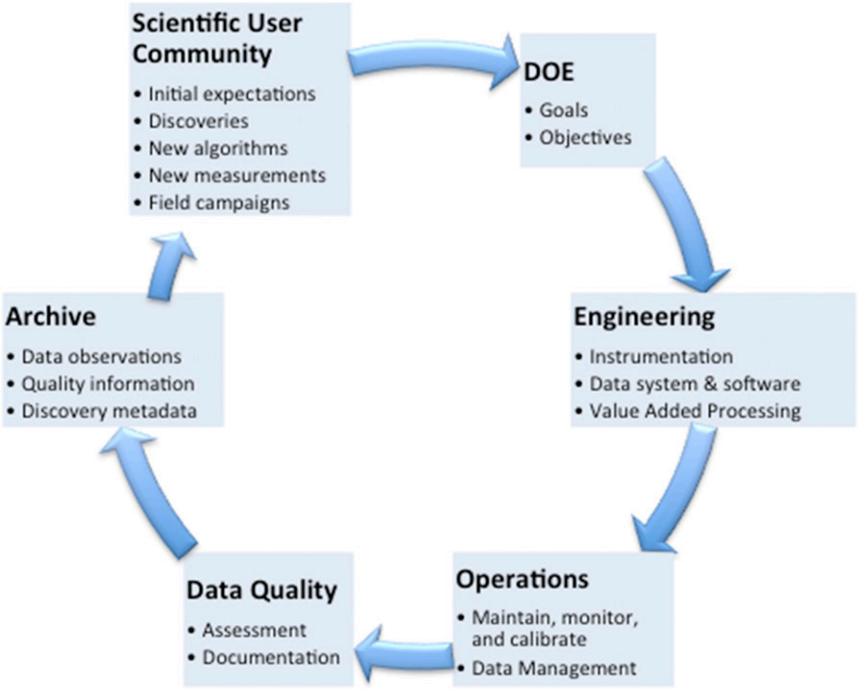

FIG. 11-2. ARM data life cycle includes programmatic, infrastructure, and external elements.

- ARM was expected to be a geographically distributed program of field sites, data processing systems, infrastructure staff, and research users.
- High-speed and long-distance networks were relatively new and had limited stability. The software for transferring data was primitive and the user community had limited experience.
- The data flow needed to be timely to keep the systems stable and to provide a reasonable timeline for initial data quality review and problem resolution.

4. The early period: 1992–97

During this early operational period from 1992 to 1997, the scope for ARM expanded from a few instruments in a single location to several facilities at a single site [Southern Great Plains (SGP)] to a second facility in operation [tropical western Pacific (TWP)] with a third facility [North Slope of Alaska (NSA)] under implementation. The data system was pushed through several changes in this period. The transition from planning to reality (real instruments in operation with output data and a user community) caused many changes. Other changes occurred as ARM implemented different system designs and data products for each of the sites. Advances in information technology also enabled significant changes during the period. The primary changes away from the original requirements for the data system during this period are summarized in the following sections.

a. The data stream

When ARM started, there was hope that a single unified information system or database could be used to handle all of the data collected by the ARM Program. However, it was quickly realized that the variety of the instruments and sites resulted in a very diverse set of datasets. This diversity made it extremely difficult to store all of the data in a single database or to use a single unified information system. Thus, the program needed another approach.

Ultimately, the basic structure of ARM data became data streams of files, where a data stream was a set of files that came from a single source (e.g., an instrument or algorithm) with a fixed file structure. The data stream became a fundamental aspect of the ARM data structure. Nearly all data stream files contained measurements spanning a single day from a single instrument and location (e.g., from a surface meteorological station at the NSA site).

b. Adoption of NetCDF

ARM, like most scientific programs, understood the need to have metadata and documentation reliably associated with its data. For example, it is important to document the SI unit and minimum and maximum

TABLE 11-1. Early ARM requirements: data life cycle.

Requirement	Explanation
Continuous data	This research expected to compare measurements with the results of models or use measurements for input to models. The plan for a continuous span of ARM measurements was a significant contrast to the dominant history of field campaigns for other observational programs.
Seamless data	The collection of measurements should approach the same completeness as model results (every value for every time step). Minimized data gaps also ensure the cooccurrence of multiple measurements needed for systems research (e.g., ''closure'' in radiative transfer).
Cumulative data	All data collected by ARM were expected to have long-term value. A cumulative view of data accommodates research on interannual variations and long-term assessment of model/data comparisons. The ARM data system should provide cumulative data retention and access for the users. Cumulative data schemes were unusual when ARM was starting.
10 yr of measurements	The minimum time range for ARM observations was proposed to be 10 yr. This duration was needed to obtain a range of climatologically significant conditions for each location. Yet, this duration was known to span multiple generations of technology.
Decades of data use	Research on models and their revisions was expected to be a long-term process. The ARM collection was also anticipated to be valuable in numerous unforeseen ways.
Data quality	ARM data were expected to have ''known and reasonable'' quality. All data should be screened with quality control limits. Intercomparisons should be made when possible. The proposed instrument design included redundancy of measurements. This provided robustness of operations and a better understanding of data quality. These quality processes should be ''near term'' to minimize duration of quality issues and insure that all data users were also provided quality information (see Peppler et al. 2016, chapter 12). The data system should provide resources for these data reviews.
VAPs	VAP processing was proposed for three primary purposes: • combining data from multiple data streams into a single integrated product (intended to facilitate easy use by researchers); • data quality intercomparisons between similar or redundant measurements (also called QMEs); and • derivation of physical quantities that were not directly measured by the instruments (e.g., cloud base height from the profiles of micropulse lidar backscatter). VAPs were expected to have a wide and ongoing timeline for development. Some VAPs were clearly anticipated. Others were expected to result from future research. The data system should provide the software development and processing capabilities for VAP creation.

values for each geophysical variable in the data stream. As a consequence, ARM adopted the NetCDF file structure because it was a self-describing, machine-independent, and efficient binary data format with an extensive set of open-source software libraries for data access and file manipulation (Rew and Davis 1990; http://www.unidata.ucar.edu/software/netcdf/). A self-describing data format was very important because ARM data files would be moved and restructured in many ways during research use. This format allowed for global metadata about the entire data file (the date of creation, source of the information, programmatic citations, etc.) and metadata for specific data fields (descriptive strings about each field, valid minimums and maximums, precision, etc.). Machine independence was important because the research community used a variety of hardware (UNIX machines, PCs running windows, etc.) and software. Furthermore, binary data formats are more efficient from a storage perspective (i.e., the binary files are smaller than ASCII files).

ARM data were intended to be widely available, and open-source software, such as the NetCDF package, insured the ability to maintain access to the data without proprietary restrictions and prevented the software from becoming abandoned because of shifting corporate interests. The NetCDF library included routines that enabled users to directly access data fields by name with their own software without knowing detailed information about file layout and field position. Tools that came with the NetCDF software distribution readily enabled users to concatenate files across time for a single data stream and retain only a portion of the data fields (to reduce the number of data files and the data volume).

The decision to select NetCDF was easy because the data format was being developed specifically to support atmospheric research. NetCDF readily supported the time, spectral, and height dimensions needed for ARM data, thereby supporting the wide variety of data structures that were stored in NetCDF format. Data streams range from simple constructs such as simple time series of radiometric data to complex multidimensional arrays from cloud radars.

The selection of NetCDF had many of the positive features listed above. However, this decision also had negative impacts on the ARM user community. NetCDF libraries required the user to write software in a programming language (e.g., C, C++, and

TABLE 11-2. Early ARM data system requirements: data heterogeneity.

Requirement	Explanation
Homogeneous structures	ARM data would be used in numerous combinations. The data format was required to be common across all of the measurements. The data structure was initially proposed to be a database and/or files with a common format.
External data	The data system should import data from other programs (NOAA/NASA satellites and weather forecast operations). These data would be used as input for VAPS and would be essential to ARM research. Processes to acquire and incorporate external data would be diverse. External data should be restructured into an ARM data format.
Open access	Accessibility should follow U.S. Global Change Research Program policies for open data sharing and maximize the use of ARM data (U.S. Global Change Research Program 1991). The user community should extend beyond ARM researchers.
Very large volumes	Remote sensing (both uplooking radars and lidars and downlooking satellites) and spectral data would have very large data volumes. In addition to the large data volumes, ARM data would also include very large numbers of small data files (e.g., surface meteorology). The design should include a mixture of online and offline storage. The cumulative volume would be much larger than most environmental data collections.
Routine and field campaign data	The data system should be able to process both files collected routinely (daily for many years) and for special studies (field campaigns and IOPs).

FORTRAN) to access and process the data, which was a significant challenge in the early days of ARM because NetCDF was a newly developed package. Not many in the atmospheric scientific community were familiar with it. Higher-level languages like Perl, Java, R, Python, MATLAB, and IDL with routines to access/read/write NetCDF files became more available in the mid-to-late 1990s, making it easier for scientists to use ARM data.

Each of these data access and processing methods required an investment in training, software development or software purchase to use the data. DOE management was initially concerned that the selection of NetCDF might restrict the use of ARM data by

TABLE 11-3. Early ARM data system requirements: design strategy.

Requirement	Explanation
Data delivery	Initial plans for ARM included coupled operation of field sites and models with quasi-real-time intercomparisons for ARM research. Field campaigns were expected to need near-term data processing for forecasting. Data delivery should be timely and uninterrupted. Data transfer to users would use high-speed networks (even before the Internet as we know it today). When network capacity was insufficient, the data transfer would include express-shipped storage media. Delays in data flow could potentially delay discovery of data quality and operational problems.
Incremental implementation	The initial design of the data system focused on rapid implementation. The data system should be functional by the start of field operations and should provide immediate turnaround for data quality assessment. The system design should use existing technology when possible.
Independent software components	Because of the wide disparity in data volumes between the data products, the software modules for data processing should be designed for independent development and operations (between sites and instruments).
Flexibility and scalability	The needs of ARM would evolve within the scope of the planned 10 yr of operations. Also, the span of ARM would exceed the hardware life cycle for data systems. Independent systems and processes were needed to assure that the data system could easily grow with development of new instruments (a major initial component of ARM).
Geographically distributed	The comparison of observations and models would have a global scope. Initial locations for field sites would be isolated by thousands of kilometers. The data system was distributed to enable parallel development of the components. Each site system should have independent operations to prevent large impacts from problems at a single location. The development of the data system was distributed across five national laboratories. The data system design accommodated the geographic spread of field sites, data systems, and data system developers.

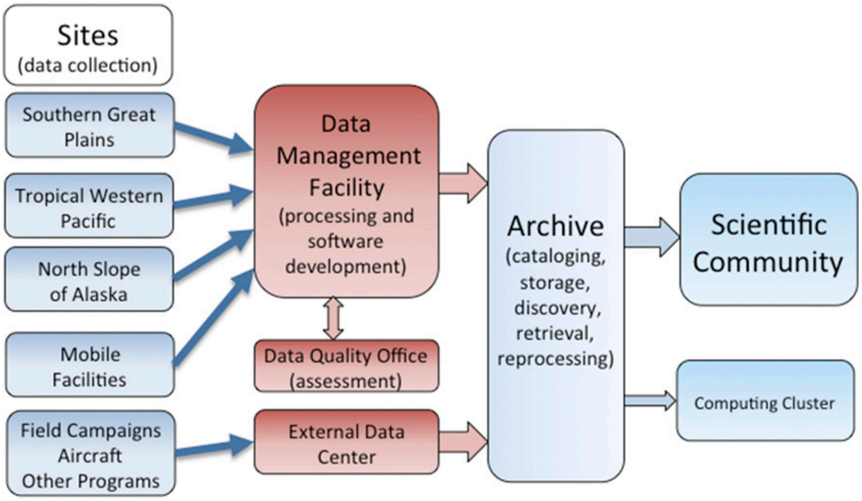

FIG. 11-3. ARM data system overview.

requiring the scientists to have programming skills to access the data, thereby limiting data use by scientists who could only use interactive tools like spreadsheets or text editors to ''look at'' data or create simple data plots. Thus, in the early 1990s, an experiment support team, which was a small group of data system staff, would work with ARM scientists to help them learn to manipulate ARM data in NetCDF files. This group was later called the ARM Experiment Center and also gathered requirements for commonly needed advanced data products. These efforts were the origin of ARM's value-added products (VAPs; Ackerman et al. 2016, chapter 3). As newer tools became available and more scientists learned how to use NetCDF, this group was no longer needed to support ARM scientists, but efforts to design and create VAPs continued.

In retrospect, the decision to organize the data into daily NetCDF files and data streams as the basic elements for the ARM data structure has been a very robust and durable decision.

c. *Raw versus processed data*

There was an axiom that was popular in the early days of the ARM Program that is typically attributed to Gerry Stokes, the first chief scientist of the ARM Program: ''one can survive bad analysis but cannot survive bad data'' (G. Stokes 1996, personal communication). This idea resulted in the decision to save the raw output from each instrument in the ARM Data Archive as well as the more highly processed (i.e., calibrated, value added) data. While this approach consumed much more space in the ARM Data Archive, it also allowed problems in both simple ingest codes (i.e., processing codes that converted raw data to NetCDF) and VAP codes to be fixed. While this might seem obvious now, it was less so when the price of storage was markedly higher and the rate of data storage much slower than today.

d. *Real-time to near-real-time availability*

The goal of autonomous 24/7/365 operations required new thinking for making the data available to the scientific community; as Stokes (2016, chapter 2) indicated, there would be no respite at the end of the campaign for which to process/calibrate/quality control (QC) the data. Thus, the program's original intent was to make the data available in real time. However, because of the relative infancy of the Internet and processing systems, data collection and processing took a finite amount of time. For a site that was well connected to the Internet such as the SGP, data could be made available to the scientific community within 2 days; however, for remote sites such as the TWP, the data would not be available for potentially several months because it was shipped back from the remote sites via transportable media (like tapes or portable disks).

The goal of making the data available in real time or near–real time was considered a positive paradigm shift in the atmospheric community, and relaxing this goal by making the data available a few days after receiving it at the primary data system was initially seen by some in the ARM community as not meeting the goal. But over the first 5 yr of the program, the scientific usage pattern of ARM data shifted from near-term comparison with models to an evaluation of a continuous series of case studies. These case studies were usually a series of intermittent days that were specifically suited for the evaluation of a common scientific question. The scientific users typically needed to access a period of data spanning a few weeks to several months. The immediate need to access recent data was generally irrelevant.

So while the scientific need for near-real-time data availability became lower priority, the program maintained the "available within 2 days" approach. This near-real-time approach allowed the infrastructure to identify instrument failures, processing errors, measurement bias, and other data quality issues more quickly (Peppler et al. 2016, chapter 12). This short delay assured that researchers were provided a stable data product with known quality.

e. Data quality

One of the program's mantras is that "ARM data have known and reasonable quality" (Gracio et al. 1996). Prior to ARM, atmospheric data were typically collected in relatively short duration field experiments, and instrument investigators would have several months after the field campaign to QC their data. ARM's operational paradigm prohibited this because there was no respite in the data collection for QC. Thus, the program had to develop automated approaches to apply the QC. Three distinctly different approaches were (are) used.

The simplest form of QC applies a set of logical rules based upon the specific data stream and its origin (e.g., the surface radiation measurements at SGP) during the data processing to set a QC flag in the NetCDF data stream file. These QC flags identified values that were too low, too high, too erratic, or failed other expectations.

A higher level of QC could be applied by comparing observations from different sources or by comparing observations to a model simulation where the model was driven by a different observed dataset. The quality measurement experiments (QMEs) were very popular in the ARM Program in the mid-1990s, with many of them developed to help evaluate radiometric instrument data quality (Mlawer and Turner 2016, chapter 14; Peppler et al. 2016, chapter 12). QMEs were automated in the data system, producing another data stream that was analyzed by scientific users and could find more subtle problems with the measurements than could be found using more simple minimum/maximum/delta checks.

Often, the best description of data quality came in the form of manually entered data quality reports (DQRs). These DQRs are more dynamic than QC flags and contain a description of an "event" that altered the normal quality of the data. Events include a wide range of conditions such as instrument degradation and contamination and temporary operating conditions (power failures, frozen or snow covered sensors, etc.). The quality impacts from these events are defined for a specific time range, list of data products, and specific measurements. DQRs are typically submitted by either the instrument mentor [see Cress and Sisterson (2016, chapter 5) for a description] or the Data Quality Office (Peppler et al. 2016, chapter 12). DQRs are now provided as "companion" information with the data files when researchers request data from the archive.

f. Pushing versus pulling data

The initial design of ARM's data flow expected researchers to define relatively static combinations of data products to be delivered to them. When data files were ready for distribution, the ARM Experiment Center would "push" the data files to the researcher's system, either over the Internet (via a program like FTP) or by copying data onto computer media and shipping these to the researcher. The "pushing of data" was implemented to meet the initial requirement for near-term comparison between measurements and modeling results and to ensure that the scientific users were getting ARM data. However, the pattern of data use quickly evolved to retrospective analyses of special case study days that occurred during the history of data collection and were relevant to the researcher's needs. Pushing a steady stream of data to researchers was not necessary for this pattern of use, and this role of the Experiment Center was discontinued. Furthermore, other advancements in Internet accessibility and web applications (see below) made it easier for scientists to select the data they wanted and to "pull" it to their machines. The flow of data to the researchers evolved from a more labor-intensive push to a more efficient and selective pull of data. With the pull method, users only worked with the required data instead of all of it.

g. The growing web and its applications

The rapidly advancing field of web applications, search engines, and the continuing development of the Internet provided new opportunities for the ARM data system to evolve. User interfaces at the ARM Data Archive allowed users to specify dynamic selections for their data needs. These selections created customized lists of data products and data ranges that resulted in data requests from the archive. The archive retrieved the requested data files from the mass storage and put them online for the researchers to download during the next week. Dynamic selections of data from the archive were more adaptable to the evolving needs of the researchers and less labor intensive for the data system staff.

The initial user interface for accessing the data inventory information of the ARM Data Archive required server class hardware and the use of X windows display capabilities on a researcher's computer. This generally required computers, network capacity, and expertise more substantial than a typical PC and scientist in the early 1990s. The software to interactively display data availability and forms to capture user data requests was tedious to develop and had limited scalability.

The invention of Hypertext Markup Language (HTML) in the early 1990s allowed an unlimited structure of text, pictures, and drawings. The addition of HTML forms enabled users to enter selection criteria into web pages. Developing software to display dynamic content on subsequent web pages based on the user's criteria was a major boon to ARM. This enabled ARM to communicate to the user manageable portions of its instrument documentation, operations records, data inventory, and data quality information. Web applications and related data retrieval processes enabled more automated processes and allowed for the initial scalability of the ARM Data Archive.

Soon after web technology became available, displaying simple lists or tables summarizing ARM's scope became more content than most users would absorb. Dynamic web displays of information were very important as the types of instruments moved into tens, the types of data products moved into hundreds, the different measurements across several locations became greater than 1000, and ARM accumulated hundreds of thousands of user accessible files.

The unlimited linkage of web-based information was important for documenting nonroutine operations from the intensive operational period (IOP) and the resulting nonroutine data products. Web links could easily accommodate the layers of detail needed to describe complex combinations of instruments and data products associated with field campaigns.

The web pages were an immediate advantage to ARM and greatly facilitated the creation of large quantities of information. However, they quickly became opportunities for inconsistent organization of information and confusion for ARM users. The ease of creating web information in common word processor software resulted in many people creating content in many different styles with inconsistent keywords and organization structure. In response to this complexity, a staff member was assigned to coordinate web-based information.

The inconsistencies shown in the web information revealed inconsistencies during many steps across the data life cycle. Many other factors led ARM into a period of improved organizational structure (or reorganization). The management, structure, and logic of the data system were also soon reorganized.

h. *Dropping operational models from ARM's scope*

In the earliest days, the program had planned on running spectral radiation and single-column models (SCMs) as part of the data system's infrastructure. However, it quickly became apparent that the challenges of collecting and managing the field measurements required all of the available resources, and thus running SCMs was dropped because the level of effort was too high. However, the spectral radiative transfer models were still run as part of the ongoing QME effort.

5. Reorganization and growth: 1998–2007

The experiences gained in the first epoch of the ARM data system highlighted many things that worked well and some that did not. In particular, many aspects of the data system were separated in ways that were not very efficient or made little sense. During this period from 1998 to 2007, the ARM Program continued to expand, and ARM managers addressed issues that arose from this growth and design decisions made in the early period.

a. *Consolidation*

The early data system management structure of ARM was site-centric, and each field site started independent and parallel implementations. This structure was efficient for getting started, but also resulted in independent designs by each site for data systems, data quality review, and data products. The implementation of multiple solutions to similar problems limited the data system's ability to keep up with the required growth needed for new sites and instruments. The data systems were a venue where many unnecessary complexities and inconsistencies became apparent from divergent evolution of site data systems and processes. As ARM accumulated more data history than the normal 3–5-yr project, continuing to support the resulting complexity was difficult. Presenting the data collection to the researchers with the archive user interface was problematic.

An infrastructure review committee recommended in 2000 that the overall structure of ARM be changed from site-centric to a structure emphasizing science-centric themes across the array of ARM measurements (ARM 2001). The revised ARM infrastructure included groups with thematic responsibilities (technical coordination, publications and communications, operations, engineering, data quality, data processing, etc.). The new programmatic structure focused common processes for developing site data systems, ARM software, data quality review procedures, and data products across all of the locations. Common solutions enabled the data system to respond more readily to continued growth from new science themes. These changes were applied not only to new instruments and locations, but also to significant redesign of existing processes and products.

Variations across the ARM data systems should now only occur to accommodate the environmental and logistical differences between the sites.

Several changes in the organizational structure of ARM impacted the data system. The production-scale processing of ARM data from all sites and VAPs, which was spread over many different locations, was concentrated into a single data management facility (DMF). The DMF succeeded the Experiment Center as a centralized location for ARM software development. The DMF component of the data system also enabled the development of extensive tools for tracking data processing and data flow. The ARM DSView tool provided numerous interrelated reports on processing status and was implemented in 2001 and updated in 2007 and 2011 (Macduff et al. 2007; ARM 2011). The same tool was also used to monitor data system operations at the sites. Data processing at the field sites was limited to the needs for operational oversight and short-term needs for field campaigns. The site data systems and instrument computers were also standardized as they were replaced during this period of ARM history.

The Data Quality Office was established in 2000 to review all ARM data and to address concerns about inconsistent data quality (Peppler et al. 2016, chapter 12), which led to several changes in the management of data quality information. A major update in most of data products in spring 2001 included a standardized structure for QC flags for most measurements. These flags included the results of tests for minimum, maximum, offset limits, and other tests specified by the instrument mentors. The DQRs were expanded to identify specific measurements and data products that were associated with a data quality event. The DQR distribution was restructured from plain text into portable web pages that included a table of contents showing the DQR titles and a tabular format for the DQR details. During this same time period, a new process at the archive began distributing DQRs that retroactively described data quality issues. A customized distribution of DQRs was prepared for each researcher based on their history of requests for data files. Attempts also were made to define a ‘‘quality color’’ (red, yellow, and green) from both the subjective judgment in the DQRs and an automated algorithm based on the prevalence of QC flags. The quality color from DQRs was attainable. Enabling auto quality color with a stable, automated process and comprehendible presentation to the user was elusive.

Other changes from ARM's reorganization impacted the structure of the overall data system. Gerry Stokes, the first ARM chief scientist, predicted that ‘‘all data in the archive would need to be reprocessed seven times’’ (G. Stokes 1997, personal communication). In the early era, reprocessing was done on the individual data system that originally processed the data (e.g., the specific site data system). However, this was very inefficient and cumbersome. Thus, a reprocessing center was set up to standardize the procedures and records for the reprocessing tasks. Formalized reprocessing enabled ARM to improve data by removing inconsistencies between sites and across time for a single measurement type (e.g., surface meteorology). Reprocessing also applied updated and more robust algorithms to the data and improved the completeness of the data collection.

VAPs are an important source of ARM data. As indicated above, VAPs combine observations from different instruments, sometimes with model output, to create new data streams. Just like the instrumentation in the program, VAP algorithms need to be robust and able to run in all conditions and thus are typically programmed by staff trained in computer science to create robust, well-documented codes. In the early years, these software developers worked directly with ARM principal investigators (PIs) to develop VAPs; however, this was not efficient because often a communication gap existed between the PIs and the developers. Thus, ARM started pairing VAP software developers with ‘‘translators’’: infrastructure scientists who could help the software developers interpret the scientific algorithms for VAPs and review interim results (Ackerman et al. 2016, chapter 3). The translators regularly coordinated the scope of the VAPs and their priority to improve the consistency of the data collection. Data distribution to all researchers (both ARM scientists and the broader research community) was assigned to the ARM Data Archive. All of these changes improved the efficiency of each of these activities.

b. *New approaches to attain consistency*

The continually expanding collection of data streams and more powerful and web-based tools to display the ARM data collection revealed yet other challenges for the ARM data system. As the number of data products increased with new locations, instruments, and VAPs, it was increasingly more difficult to maintain consistency for the design of the data products. Detailed aspects of the data product design such as data field names and quality control limits were inconsistent between similar data (e.g., across all of the meteorological or radiometric measurements). This inconsistency was confusing and annoying to the data users. In 2004, efforts began to put all details for the data fields from all data products into a database. This change enabled differences between similar measurements to be readily identified and consistency easier to attain. This database became a production feature of the data system in 2006.

The continually increasing scope of ARM data made it more difficult to determine the details of the inputs required for VAPs and their availability. A database was developed to define the dependencies between the VAPs and their input data streams. This dependency information included the details about data fields needed for input into each resulting VAP data field. This dependency database was added to the data stream configuration information previously described.

Inconsistent ARM language became apparent as the ARM data products began to converge toward consistency. The size and scope of the data collection expanded rapidly in 2000–04 and also contributed to inconsistent language. For example, the thematic categories of measurements used for data discovery at the archive and for the web-based ARM documentation were similar but different enough to be confusing to the researchers. A collaborative effort in early 2006 defined and adopted a metadata structure to be used by both data discovery and web organization. After this metadata consolidation, many of the archive user interfaces and the ARM web pages could be composed dynamically from the common metadata content stored in a single database. These changes assured that ARM users encountered the same language and logic as they interacted with ARM information.

The incompleteness of the metadata was a critical problem when the display of ARM information and access to data became dependent on the common metadata. Throughout this entire period, efforts were made to encourage the creation and recording of metadata about the data products, measurements, instruments, and sites. The initial solution included educating many members of the infrastructure about the scope of metadata and its role in communicating ARM information and data to the researchers. The infrastructure staff contributing metadata ranged from instrument mentors to software developers and translators to data quality staff. The next step included formally defining the metadata elements and developing templates to record and transmit the metadata. Establishing informal processes to referee the consistency of the metadata was the final step. Timely creation of metadata was an elusive objective. ARM shared this challenge with most other research programs during this era.

During this midportion of ARM's history, the number of data products from field campaigns grew beyond a short list that could be shown on a single web page. The complexity of the data collection from field campaigns was challenging. The structure and scope of field campaign data varied with each campaign. Standardization across the field campaign data was limited because the collaborating researchers used specialized methods for processing and organizing the data. Standardizing the data structures required too many resources. Evolving standards for data structures also limited standardization. Most of the field campaign data followed a consistent structure for each campaign and instrument. More detailed data structures were unconstrained. To accommodate this complexity and retain a maximum amount of these data, the field campaign part of the data system was designed with the following principles:

- When requested, researchers were provided guidance on file names and directory structures.
- A simple FTP procedure or shipped media was used for data transfer from the researchers to ARM.
- The data were organized in a master directory structured by year, site, campaign, instrument, and PI name.
- Documentation for each subdirectory was requested to explain the scope of additional data directories and filenames.
- A web-based user interface was developed that could navigate and provide access to any amount of documentation, complexity, or organizational depth for the data collection.

These design principles were found to be robust as the field campaign data collection grew. Both data providers and data users readily accepted this design.

c. *Communicating the scope of ARM's data collection*

The increased size and complexity of the data collection during this period also challenged the researcher's ability to understand what data were available from the ARM Program over its duration. Simple query logic that searched for the availability of data based on criteria of location, measurement type, and date range failed to find data because the incremental implementation of sites, instruments, and VAPs. The failed queries provided no information about why data were not found (the time range was invalid for a specific location, instrument was offline for repair, etc.). Additional interfaces were developed for the archive that presented hierarchical summaries or catalogs of data availability (McCord et al. 1999).

The expanding data collection also challenged the researcher's ability to understand the ARM results. Comprehension was limited because scanning the contents of the data files was not possible. Web-based displays showing plots of ARM measurements were developed. In 2004, a user interface (thumbnail browser) went into production at the archive that displayed tens of thumbnails (standardized, postage stamp size, and data plots) per web page. The data plots (both small thumbnails and labeled page-size graphs) were precomputed to assure a responsive and predictable performance of this user interface as the size of the ARM dataset expanded. This user interface allowed researchers to customize displays of specific measurements from multiple data products and visually scan the results in monthly increments (ARM 2004).

A web-based tool for interactively plotting small amounts of NetCDF data (NCVweb) was also developed during the same time period (Bottone and Moore 2003). This NCVweb tool allowed users to select relatively small time ranges of data and ''zoom in'' on the details of individual measurements. The NCVweb tool was expanded to display the details of the data fields and export small portions of data into text files or smaller NetCDF files. Both the NCVweb and thumbnail browser user interfaces for the archive are still used by researchers. Routinely generated data plots and interactive visualization of data with NCVweb also are used during data quality review [see Peppler et al. (2016, chapter 12) for more details].

The primary themes contained in the midperiod of ARM's data system history included

- attaining efficiency through the consolidation of data system functions;
- expanding the creation and communication of data quality information;
- converging on consistency in the detailed structure and description for ARM data products and documentation; and
- improving the communications with the researchers through new interfaces to present the scope of the data collection and visualization of ARM's results.

All of these changes continued to help ARM's success in the final era of the first 20 yr.

6. Continual growth and improving durability: 2008–present

During the period between 2008 and the present, ARM's data system had to adapt to increases to the number of new instruments and sites. Also, the designation of the ARM facility as a national scientific user facility resulted in ARM no longer being purely a research program but now becoming a truly operational facility. To sustain operations and durability during these changes, it was necessary for the data system to be more robust and nimble. The changes and continued growth of ARM's data system are summarized in the following sections.

a. Challenging growth opportunities

Persistent themes for ARM during this most recent period of ARM's history were growth and accelerating growth. These themes originated from implementing several new sites (e.g., the new permanent site in the Azores) and instruments as well as more tools for efficient data product development. The primary events that triggered growth during this period include the addition of

- a second mobile facility (2008–10);
- the purchase of many new types of instruments with Recovery Act (stimulus) funding (2009–11; Mather and Voyles 2013; Voyles and Mather 2010);
- acquisition of scanning instruments with new types of data products (2010–12);
- implementation of new fixed location in the Azores (2012–14); and
- implementation of a third mobile facility (2012–14).

Many of the Recovery Act instruments included the ability to measure profiles, three-dimensional (3D) scans, spectral views of solar radiation or Doppler shifts in radar, and lidar signals. For example, the daily data volume for the cloud radar dataset increased from 15 to 20 to 80 to 100 GB day^{-1}. Improvements in instrument, sensor, and computing technologies also contributed to rapid increases in growth rate in the overall data volume collected by ARM. All of these changes contributed to dramatic increases in the volumes of data moving daily from the sites through the DMF to the ARM Data Archive (Fig. 11-4).

The designation of ARM as a national scientific user facility in 2004 (Ackerman et al. 2016, chapter 3) converted ARM from a 10-yr project to a permanent operation for atmospheric measurements, data collection, and data distribution. Performance metrics for the facility were defined for data completeness and increasing the number of users. This designation encouraged ARM to aggressively identify scientific collaborators for mobile facility implementations and other smaller-scale field campaigns. A significant number of new users are added to the ARM user community with each new mobile facility location. All of these factors increased the diversity of data users. The longevity of ARM encouraged researchers to evaluate larger quantities of data. The rate of data usage by the researchers also accelerated (Fig. 11-5).

As the increase in the outgoing and incoming data volume ranged from 10 to 120 times respectively during this period, the data system required continuing

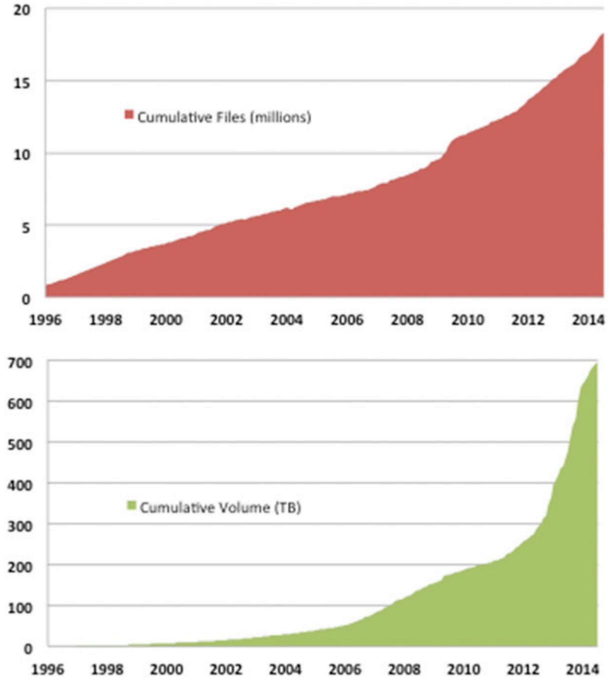

FIG. 11-4. Cumulative data volume stored in the ARM Data Archive in terms of (top) number of files (millions) and (bottom) volume (terabytes).

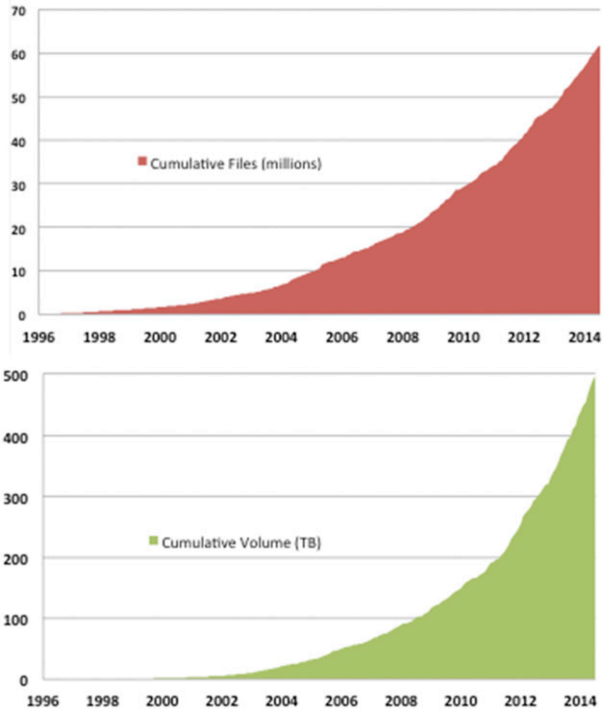

FIG. 11-5. Cumulative data volume requested from the ARM Data Archive in terms of (top) number of files (millions) and (bottom) volume (terabytes).

revisions to its design in order to remain nimble. Several changes occurred in the data system to help it reduce the complexity in its design and the chaos in its performance. The changes summarized in the remainder of this section enabled the system to more quickly adapt to the changing requirements.

b. *Expansion and coupling of hardware*

Expansion for data volume was accommodated primarily with increased data storage at each point in the data life cycle. Data access was increased with high-performance networks and high-performance file servers. The file servers concurrently used several hardware interfaces and numerous storage devices. In 2012, the DMF hardware was moved from Pacific Northwest National Laboratory, which had operated the DMF and its predecessor the Experiment Center from the beginning of the ARM Program, to Oak Ridge National Laboratory. This move enabled the DMF network and storage devices to be combined with the archive. This change enabled quicker adaption and reallocation of the combined hardware capacity to different parts of the data life cycle as needed. In an analogous manner, the site data system has been modified by consolidating individual instrument computers into virtual machines on a single redundant pair of computers with replicated storage. This design reduces maintenance efforts and provides the capability to transition processing from one set of computers to another if the first fails at any of the sites. All of these changes allow the various ebbs and surges of data flow to share common storage buffering resources that are more robust and durable than can be implemented for each process.

c. *Consolidation and software efficiency tools*

The very large number of new instruments purchased with the Recovery Act funding resulted in the need to build an integrated software development environment (ISDE) as part of the data system. The vision for ISDE included providing standardized tools for the common software tasks. The availability of ISDE would reduce the effort for new software and allow development to focus on the specialized portions of the software. ISDE was also intended to be portable so that software developed by researchers would more readily follow ARM standards for software and data products. After 3 yr of experience with ISDE, its scope was reduced to handling the data input, combining data from multiple inputs into a common interval for sampling rate, data product design specifications (from the existing database), and handling the output. Most of the software that uses ARM data requires some form of these functions. ISDE was renamed the ARM Data Integrator (ADI) to better represent the scope of its functions. ADI was valuable because it reduced the effort for developing new data products (i.e., VAPs). It also focused most of the software development effort on the scientific logic because the reading and writing, as well as the time syncing of different data streams, was now done internally by ADI. ADI also enabled researchers to develop more standardized software and data products when collaborating with ARM.

d. *Reducing the complexity*

The expansion and requirements for the durability of ARM data processes resulted in the formalization of the VAP development process. In the previous periods of ARM, VAP development could be a very lengthy series of iterations between the translators, the software developers, and the ARM PIs. Earlier styles of VAP development included incomplete processes for defining the data product, selecting advanced algorithms, evaluating the results, and declaring completion of the development. These earlier styles sometimes resulted in endless development processes for VAPs. The formalized process defined a specific series of evaluations for the following:

- confirmation of algorithms to be used and knowledge of their logic;
- decisions to release interim versions of the VAPs data for review by the researchers;
- assessment of the interim progress on the VAP development; and
- creation of documentation of the finalized data product.

The process also recognized that some VAP developments might be abandoned or left as is because further progress depended on new research beyond the scope of ARM. In some instances, these intermediate VAPs were contributed by researchers as PI data products. These PI data products enabled the research community to make additional discoveries until the methods for a final product were developed. These revisions to the VAP development process enabled ARM to meet continuing challenges from a very large number of possible VAPs associated with the expanding number of sites and instruments.

The program realized that PI data products offered an excellent way to get additional higher-order processed data into the ARM Data Archive, thus both capturing it and making it available to other scientists. The program created an area in the archive specifically for value-added datasets provided by ARM PIs. With this procedure, a translator worked with the PI to move the data into the archive. Then, a metadata editor was developed allowing the PI to include metadata about the contributed dataset that both described the algorithm used to create the dataset and provided key words so that other PIs could more easily find the dataset within the archive holdings.

The reprocessing of different datasets over time led to multiple versions of data products being used in research projects, especially for those PIs performing long-term analyses. Standards for data products were needed to reduce the complexity of the data processing by the researchers. ARM developed a standards document (and revision process) for data product names and data product design as an extension of the standardization theme in 2012/13. An infrastructure committee representing all portions of the data system, the data life cycle, and a few data users developed these standards. Sustainable standards have enabled the ARM data system to readily expand. They also facilitated the long-term durability of the data products and related software.

Sustaining standards for the names and descriptions of ARM data required improvements for creating and reviewing the metadata. Development of a web application to formally track the proposal, review, and approval of the metadata began in 2012. This application enabled ARM to improve the completeness of the metadata and expand the scope of metadata creation to include the field campaign data. Formalized metadata for these data were needed to consolidate the discovery of routine and field campaign data into a single user interface for the archive. This consolidation has improved the durability of the user interface and helped to expand the researcher's data discovery.

e. *Improving the use of very large ARM datasets*

Some of the new instruments acquired in the late 2000s as part of the Recovery Act, such as the scanning cloud radars, Doppler lidars, and aerosol mass spectrometers, create huge amounts of data every day. These files are sufficiently large (many hundreds of gigabytes per day) that transferring these files to PI computers was challenging and very inefficient. To address this, special computers were added to the ARM data system to provide the researchers a means of exploring these new and very large data. These systems (the ARM computing cluster) were implemented in late 2010. They included a cluster of processors, very large amounts of shared memory, and tens of terabytes of storage and workspace. Popular scientific software applications were installed on these systems to facilitate the analysis of these datasets, and dozens of researchers explored the large data products. This addition was another revision to the data system that supported expansion and durability.

Another strategy toward expansion and durability of the data system was making model-ready, condensed, and very specialized (showcase) data products. The development of these showcase data products started in 2007. The processing for showcase data products included new steps such as best-estimate logic that replaced missing values from a primary instrument with the next best choice of an equivalent measurement from a different instrument. Additional quality evaluation and summarization provided the best possible long-term representation of ARM's core measurements. A showcase, best-estimate product was produced for the full duration of ARM's history for all sites (Xie et al. 2010). Several data products with extensive summarization, integration, and preparation were expected to further engage the researcher community with ARM data.

The expansion and durability of ARM's data systems were assured by the consolidation of hardware, software, and development resources into configurations that supported increased flexibility. The consolidation enabled the sharing of common and standardized components. During consolidation, changes occurred to data storage, user interfaces, software development tools, showcase data products, and review processes for interim data products and metadata. All of these revisions contributed to the efficiencies needed for the data system design to adapt to the dramatic growth of ARM's measurement resources from new instruments at new locations in this most recent period of ARM's history.

7. Summary

Plans for the ARM data system began very early in the planning for the ARM Program. This planning strategy was motivated partially by extensive and complex expectations for the data system needed to accomplish ARM's research objectives. The early participants in ARM also knew that the capacity requirements for the data system were very large and complex when compared to contemporary systems supporting environmental measurements in the early 1990s. Many of the tasks in the early history of ARM were organized into parallel and incremental steps to expedite the implementation of sites, instruments, and data systems. This organizational strategy benefited ARM, but also accidentally resulted in the implementation of different designs for data systems, data review practices, and data products for each site.

The divergence of many aspects of ARM's infrastructure became problematic when the data accumulation extended beyond the duration of a normal research project life cycle (3–5 yr) and the implementation of multiple sites occurred. The divergence was especially critical because the inconsistency and complexity of the ARM data collection made research analyses more difficult. Even communicating the scope of ARM datasets became challenging.

The midperiod of ARM's data system history included significant reorganization of the ARM infrastructure and the data systems. The independent development of the data system and data products by each site was consolidated into centralized groups for many of the tasks. Centralization allowed ARM to address many of the implementation challenges with a smaller number of solutions. Many of the components of ARM's information about configuration, quality, and descriptions were migrated into shared databases for use in many ARM functions. This consolidation of information provided a consistent view of ARM documentation and data to the researchers. These databases also enabled the development of dynamic (customized) displays of information so that researchers could interact with smaller portions of the ARM documentation and data collection. Visualization of ARM's measurements were created for large portions of the data collection and made accessible on the Internet. The researchers could more quickly scan and understand ARM's results.

In the final period of ARM's data system history, the data system adapted to the very large jumps in the number of new instruments and sites. Furthermore, the designation of the ARM facility as a national user facility resulted in a change of thinking; ARM was no longer purely a research program but had to be concerned with running as a truly operational facility. To sustain operations and durability during these changes, it was necessary for the data system to be more robust and nimble. These characteristics of the data system were attained by revisions that enabled more sharing of computing resources between the DMF and ARM Data Archive. Streamlined management and configuration of site data systems resulted from the implementation of virtual machine technology. Establishing standards for the naming and design of the ARM data products reduced the growth of the data complexity. Formalized processes for VAP software development and review of metadata improved the efficiency of these efforts. Implementation of shared software tools improved the development of ARM software. Expanded computing resources dedicated to the analysis of data without download enabled the researchers to effectively use ARM data and keep up with its rapid growth. The development of highly integrated and best-estimate data products facilitated the use of ARM data by a broader user community.

Acknowledgments. The authors thank the DOE ARM Program for supporting their effort to create this publication and for their many interesting opportunities and experiences while working with ARM. Dave Turner is also thanked for his encouragement and suggestions during the revision and finalization of this publication. Thanks also to the numerous and talented ARM staff members who have created and operated the ARM data system throughout its history.

REFERENCES

Ackerman, T. P., T. S. Cress, W. R. Ferrell, J. H. Mather, and D. D. Turner, 2016: The programmatic maturation of the ARM Program. *The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years, Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., doi:10.1175/AMSMONOGRAPHS-D-15-0054.1.

ARM, 2001: The Atmospheric Radiation Measurement Program Infrastructure Review Report (AIR): Summary of recommendations. U.S. DOE Rep. DOE/SC-ARM-0001, 3 pp. [Available online at http://www.arm.gov/publications/programdocs/doe-sc-arm-0001.pdf.]

——, 2004: New thumbnail browser for data archive gives users more options. ARM Climate Research Facility News, accessed 13 August 2014. [Available online at http://www.arm.gov/news/facility/post/1109.]

——, 2011: Welcome to the matrix: Overhaul of data system status viewer hits bull's-eye. ARM Climate Research Facility News, accessed 13 August 2014. [Available online at http://www.arm.gov/news/facility/post/14503.]

——, 2016: Appendix A: Executive summary: Atmospheric Radiation Measurement Program Plan. *The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years, Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., doi:10.1175/AMSMONOGRAPHS-D-15-0036.1.

Bottone, S., and S. Moore, 2003: Tools for viewing and quality checking ARM data. *Proc. 13th Atmospheric Radiation Measurement (ARM) Science Team Meeting*, Broomfield, CO, ARM, 1–12. [Available online at http://www.arm.gov/publications/proceedings/conf13/extended_abs/bottone-s.pdf.]

Cress, T. S., and D. L. Sisterson, 2016: Deploying the ARM sites and supporting infrastructure. *The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years*, *Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., doi:10.1175/AMSMONOGRAPHS-D-15-0049.1.

Gracio, D.K., and Coauthors, 1996: Data systems for science integration within the Atmospheric Radiation Measurement Program. *Proc. 12th Int. Conf. on Interactive Information Processing Systems (IIPS)*, Atlanta, GA, Amer. Meteor. Soc., 327–336.

Macduff, M., S. Choudhury, and J. Daily, 2007: The new DSView based on portal technology. *17th Atmospheric Radiation Measurement (ARM) Science Team Meeting*, Monterey, CA, ARM. [Available at http://www.arm.gov/publications/proceedings/conf17/poster/P00144.pdf.]

Mather, J. H., and J. W. Voyles, 2013: The ARM climate research facility: A review of structure and capabilities. *Bull. Amer. Meteor. Soc.*, **94**, 377–392, doi:10.1175/BAMS-D-11-00218.1.

McCord, R. A., D. J. Strickler, B. M. Horwedel, R. C. Ward, and S. W. Christensen, 1999: A catalog-based user interface enhances data selection from the ARM archive. *Proc. 15th Int. Conf. on Interactive Information and Processing Systems (IIPS)*, Amer. Meteor. Soc., Dallas, TX, 105–108.

Mlawer, E. J., and D. D. Turner, 2016: Spectral radiation measurements and analysis in the ARM Program. *The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years*, *Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., doi:10.1175/AMSMONOGRAPHS-D-15-0027.1.

Peppler, R., K. Kehoe, J. Monroe, A. Theisen, and S. Moore, 2016: The ARM data quality program. *The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years, Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., doi:10.1175/AMSMONOGRAPHS-D-15-0039.1.

Rew, R. K., and G. P. Davis, 1990: NetCDF: An interface for scientific data access. *IEEE Comput. Graphics Appl.*, **10**, 76–82, doi:10.1109/38.56302.

Stokes, G. M., 2016: Original ARM concept and launch. *The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years*, *Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., doi:10.1175/AMSMONOGRAPHS-D-15-0021.1.

U.S. Department of Energy, 1990: Atmospheric radiation measurement program plan. U.S. DOE Rep. DOE/ER-04411990, 121 pp. [Available online at http://www.arm.gov/publications/doe-er-0441.pdf.]

U.S. Global Change Research Program, 1991: Our changing planet: The FY 1992 U.S. Global Change Research Program. Committee on Earth and Environmental Sciences, 90 pp. [Available online at http://data.globalchange.gov/report/usgcrp-ocpfy1992.]

Voyles, J. W., and J. H. Mather, 2010: Recovery Act instruments: Deployment and data processing plans. *Extended Abstracts, First Science Team Meeting of the Atmospheric System Research (ASR)*, Bethesda, MD, ARM. [Available online at http://asr.science.energy.gov/meetings/stm/posters/poster_pdf/2010/P000257.pdf.]

Xie, S., and Coauthors, 2010: Clouds and more: ARM climate modeling best estimate data. *Bull. Amer. Meteor. Soc.*, **91**, 13–20, doi:10.1175/2009BAMS2891.1.