U.S. DEPARTMENT OF ENERGY | Office of Science

DOE/SC-ARM-14-008

Data Management Facility Operations Plan

NN Keck

January 2014

ARM
CLIMATE RESEARCH FACILITY

DISCLAIMER

This report was prepared as an account of work sponsored by the U.S. Government. Neither the United States 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 U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

DOE/SC-ARM-14-008

Data Management Facility Operations Plan

NN Keck

January 2014

Work supported by the U.S. Department of Energy,
Office of Science, Office of Biological and Environmental Research

Acronyms and Abbreviations

AAF	ARM Aerial Facility
AMF	ARM Mobile Facilities
ARM	Atmospheric Radiation Measurement
BCR	Baseline Change Request
CM	Configuration Management
DDT	Data Delivery Tracking
DMF	Data Management Facility
DQPR	Data Quality Problem Report
DSDB	Data System Database
FTP	File Transfer Protocol
NetCDF	Network Common Data Format
NSA	North Slope of Alaska
NTP	Network Time Protocol
ORNL	Oak Ridge National Laboratory
PNC	Package Notification Configuration
PNNL	Pacific Northwest National Laboratory
RAB	Reprocessing Advisory Board
SDS	Site Data System
SGP	Southern Great Plains
TWP	Tropical Western Pacific
VAP	Value-Added Product

Contents

Figures

1.0 Scope

This document describes Data Management Facility (DMF) Operations activities, policies, and services for the Atmospheric Radiation Measurement (ARM) program. It does not authoritatively document software or its configuration, although some details are included. It establishes a baseline of expectation within the ARM management for the DMF.

2.0 Introduction

The DMF has evolved with the ARM program around the activities of value-added product (VAP)/ingest processing, data system development, and remote site support. The ARM program has matured, and with the division of Engineering and Operations, the DMF's role formalized within Operations. While much of the remote site support has been incorporated into the Site Data System (SDS) Operations group, the processing and management of data from all sites is still central to the DMF's focus. The latest revisions of data system software have improved the automation of much of the processing. However, VAPs still require some manual effort to run reliably. As a centralized facility with experienced data operators, the DMF is ideally suited to monitor, manage, and mitigate data flow interruptions.

The operation of the DMF is composed of both system and data maintenance. These services are available to support the ARM program, Monday-Friday, during normal business hours (Pacific Time Zone). Pre-arranged weekend and after hours support are available as needed. Please see Appendix A for additional contact, delegate, and emergency information.

This document begins with a brief description of the DMF, followed by a discussion of data flow operations and a listing of general Operations responsibilities.

3.0 Description

The DMF is the data center that houses several critical ARM services including first level data processing for the ARM Mobile Facilities (AMF), North Slope of Alaska (NSA), Southern Great Plains (SGP), Tropical Western Pacific (TWP), and Eastern North Atlantic sites, VAP processing, development systems, and other network services. Currently ARM Aerial Facility (AAF) data are not processed via the DMF.

The design of the DMF is based on a simple architecture. The DMF is composed of a network with a large network file system (NFS) server, providing file service to all of the computers. This hardware is maintained by Oak Ridge National Laboratory (ORNL) staff. The data system software used on the production system is developed and supported by the ARM Engineering group and released according to the configuration management (CM) standards and the ARM baseline change request (BCR) process.

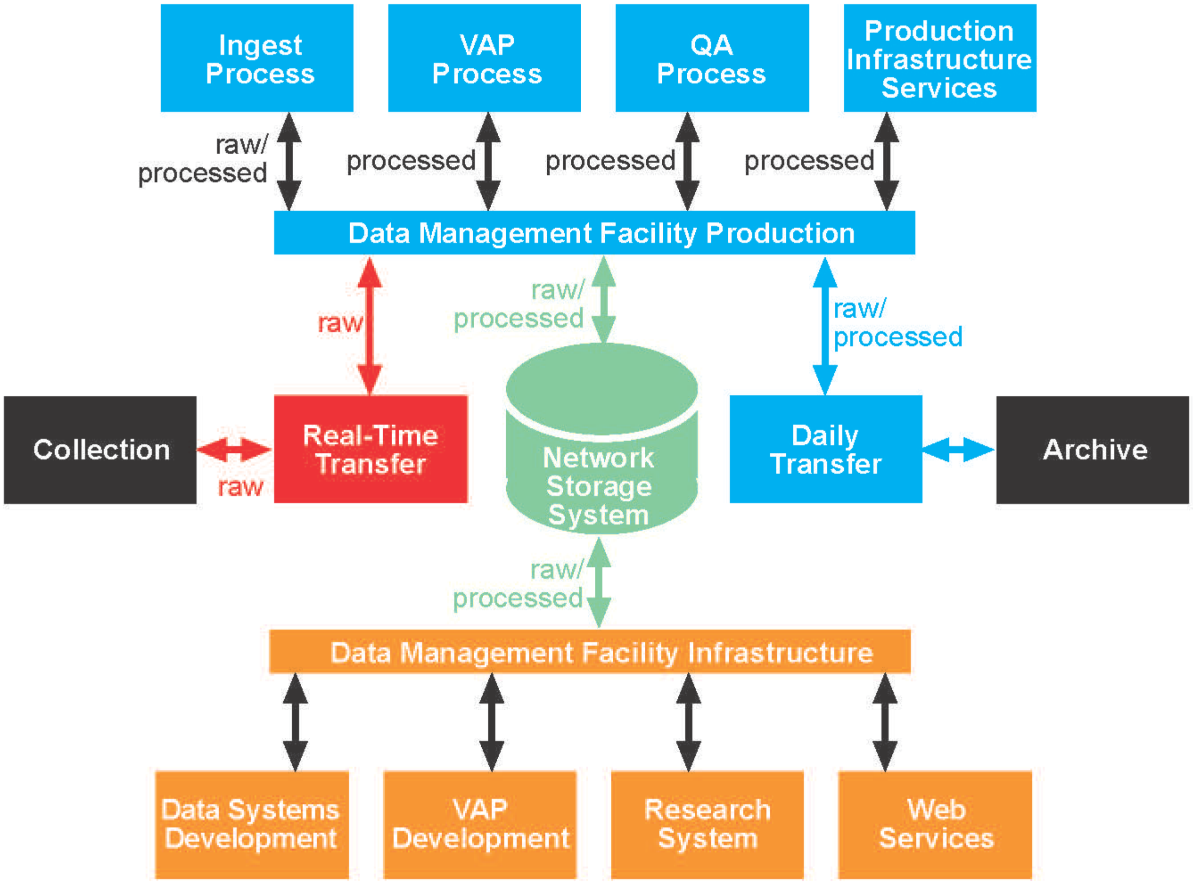

Figure 1. DMF Architecture.

3.1 Production

The DMF's production system is responsible for the official processing of all of ARM's primary observation sites' raw data. Each hour, the ARM SDSs collect raw data from each instrument and deliver it through File Transfer Protocol (FTP) to the DMF production data system. The DMF production data system runs software, called ingests, to process the raw data into the ARM standard Network Common Data Format (NetCDF) file, as designed by the instrument mentor and ARM Engineering developer. The ARM Data Quality Office works with the DMF staff to provide general quality analysis of data that is available on the DMF production system. The DMF production system also runs software, called VAP, to apply additional scientist-derived algorithms that enhance data for use by the scientific user community. All of the data on the DMF production system, raw, NetCDF, and VAP, is delivered to the ARM Data Archive for distribution to the scientific user community.

The DMF production data system software is maintained in the ARM development system software repository and deployed via the Software Package Manager, https://engineering.arm.gov/armpm/Main.html in order to provide reliable software deployment to other ARM systems. In addition, the DMF production data system also handles VAP software. Generally, the only documentation available for running VAPs is in the package notification configuration file (PNC). The PNC file contains all the metadata about the release of a particular package, and notifies interested parties that a package has been released, along with providing those parties with pertinent information related to the software. It is incumbent on the DMF Operations staff to read and understand the PNC file that comes with each VAP.

The DMF Operations staff runs, monitors, and configures multiple configuration files (Zebra, dscron, transfer) that aid in 24/7 archival and data processing. The DSView (https://dsview.arm.gov/Main.html) is a graphical user interface tool that displays the status of processes monitored by the DMF. These processes include: instrument time synchronization checks, data collections, data ingests, data rename, data bundle, and data archival. Additionally, DSView for VAPs (https://dsview.arm.gov/vap/Main.html) is a graphical user interface tool that displays the status of those VAP processes monitored by the DMF. These processes include: VAP processing, summary quality check (QC) processing, and VAP data archival.

The DMF production file server currently provides historical copies of local data for a set period of time, while data replication is completed through a separate system offsite. The ARM Data Archive is responsible for maintaining an official copy of all ARM data.

3.2 Development

The DMF development system is used by many ARM developers to design, write, and test software. The DMF development system also hosts the ARM software repository, which is used to provide reliable software deployment to other ARM SDSs.

ARM developers rely on many software packages and key services to be available on the development system. Fundamental packages related to data system utilities, databases, file modification, graphics, plotting, debugging, and compiling are a necessity. Many developers also rely on the web server as a production service, which requires the license managers to be configured and running. The ORNL system administrators ensure these important services are operational.

3.3 Other

Product Delivery web servers are maintained by ORNL. However, the DMF maintains the content of www.dmf.arm.gov. The ORNL system administrators ensure important services are operational.

4.0 Data Flow Operations

This section will describe the basics of “how” data flows hourly and daily to and from the DMF. Due to the complexity of this process, there are many steps that require monitoring. Data flow phases are categorized as Online, Processed, Stored, and Available. They are reported in near-real-time on the Data Delivery Tracking tool (https://engineering.arm.gov/ddtrack/#v::lo).

4.1 Online Phase

Using the ARM standard site_transfer, the DMF ships and receives data throughout each day. The DMF Operations monitors this closely to ensure expected data flow from all of the ARM site sources, the External Data Center (XDC), and ultimately, to the ARM Data Archive.

Health and Status packets are sent hourly from each ARM site. These packets are used to populate the DSView and Data Delivery Tracking webpages, which reflect the state of an instrument's collection, ingest, and transfer status. These packets are received with site_transfer and put away with a transfer process.

The DMF receives raw data hourly from the SGP, NSA, ENA, and TWP sites and from AMF sites when possible. Data are also sent on a regular schedule from the XDC site, contributing additional input datastreams for certain VAP processes. All of the sites' data are delivered to the DMF either through the site_transfer process, or external hard drive for larger data sets, then automatically stored in the data tree by an internal transfer process.

In addition to data collection during the Online phase, instrument time is also verified against a Network Time Protocol (NTP) server and relayed. If the instrument's time is +/- 3 seconds from the referenced NTP server, it is reported to operators via email and logged in the appropriate communication log. The instruments' time is then adjusted back to the correct setting by one of two mechanisms; ARM specifically designed data collection software, or NTP software, depending on the particular instrument being addressed.

The DMF Operations identifies anomalies and coordinates short- and long-term resolution with other Operations groups.

4.2 Processed Phase

There are several types of data processing performed at the DMF. Most raw data from the ARM sites are shipped to the DMF and processed hourly to provide near-real-time data for ARM site scientists, mentors, and the Data Quality Office. The processed instrument data are also used as input to many of the VAPs that are run at the DMF. Minor near-real-time data reprocessing also occurs at the DMF.

4.2.1 Processing Raw Data

The transformation process that converts the raw data collected from the instruments into standard ARM format is referred to as the "ingesting" of data. The application that performs the transformation for a specific instrument is referred to as the instrument's "ingest". Ingests are scheduled to run via dscron and generate data sets in a format acceptable for representing ARM scientific data, which is called NetCDF.

The status of each ingest's processing is sent by email to the DMF operator as events occur, and also is updated through the DSView and Data Delivery Tracking tools when a state changes. The DMF operator reviews the logs for error messages and responds appropriately.

Calibration information files are sometimes necessary for select instrument types. The assigned instrument mentor provides the appropriate information to the DMF through the DMF doorstep, FTP, or email.

Raw data sets are checked for continuity. If gaps are identified, the DMF operator begins to investigate the cause. The instrument may be broken, out for repair, not properly collecting data, etc. Occasionally, the data are available and just need to be sent to the DMF. When an instrument or data concern arises, the DMF operator communicates with the SDS operators, Data Quality Office, instrument mentor, and then through established tracking tools such as the Data Quality Problem Report (DQPR) to further track down and communicate the problem.

Ingested data are available for data QCs, which are done by the Data Quality Office through tools running on the DMF.

In operating the DMF, some working knowledge of the Data System Database (DSDB) is required. The DSDB is used by ARM as part of the ARM SDS to collect, transform, and manage the flow of scientific data from the source of origin to the final repository. The scope of information stored in the DSDB includes a list of locations, a list of processes, the current state and status of these processes and the datastreams produced by these processes. Understanding how processes use the DSDB can be a great help in troubleshooting and assisting engineering in resolving processing problems quickly.

4.2.2 Modifying and Deleting Raw Data

Sometimes raw data arrives with minor abnormalities, such as a duplicate line of identical data, or missing header information. This needs to be adjusted in order for the ARM ingest to process it. A mentor/developer will be contacted when bad raw data are received by the DMF that does not fit standard troubleshooting protocol. A DQPR may also be opened to track further problems with an instrument. If data cannot be fixed, it will be renamed by removing “raw” and replacing it with “bad” in the file name, so it can be distinguished at a later time. It will then be bundled with the raw tar bundle for the day it was associated with, and later shipped unprocessed to the ARM Data Archive.

Bad Raw and NetCDF files will be deleted from the DMF under the following conditions:

- an authoritative source requests the removal of data through an Engineering Work Order or Data Quality Report (usually the DMF operator and reprocessing coordinator handle such requests)
- the original raw or processed data has been, or will be, archived
- empty and zero byte files will be discarded.

4.2.3 Value-Added Product

A VAP is an additional level of processing, which creates many higher data level products that are available to users. VAPs are algorithms that use one or more data streams (instrument or VAP) as input and create one or more data streams as output. VAPs can be simple averaging routines, qualitative comparisons, or complicated algorithms for calculating required experimental data that cannot be directly measured via instrumentation.

VAPs are run daily, weekly, monthly, or whenever the required input data becomes available. The VAPs PNC file is maintained to reflect processing requirements to ensure the desired output is generated. Before a VAP can be run, data availability and continuity must be verified. A working knowledge of data streams and processes is necessary to determine when and how a VAP can be run. There are roughly 45 unique VAPs running at the DMF, with many more under development.

The key effort to producing VAP data is managing input data streams and monitoring process logs for problems, working with developers to resolve problems. Interpreting the varied output of the many VAPs requires a certain level of experience. DMF Operations has developed this expertise to ensure efficient VAP processing. The DMF has also established a relationship with the developers to quickly troubleshoot and resolve any problems encountered. Efforts to help automate VAP processing are currently being implemented.

VAP data are available for data QCs, which are done by the Data Quality Office. VAP Quicklooks are currently retained indefinitely and available for viewing at: http://c1.dmf.arm.gov/data/process/vap/calendar/ql.php.

4.2.4 Reprocessing

Reprocessing can either be ingest or VAP related. A data set may have to be reprocessed because a variable may have changed or a better algorithm was developed to interpret the data. Reprocessing data is identical to the original processing but for a specific time period. Reprocessing requests are prioritized based on translator and or program requirements.

Reprocessing efforts are coordinated with the Reprocessing Advisory Board (RAB) and shared with the Reprocessing Center. Minor real-near-time reprocessing is done by DMF Operations on normal processing systems. Some reprocessing efforts are unique enough that a developer must manage the reprocessing. In this case, the DMF provides the computing resources and ensures the data are delivered to the ARM Data Archive.

Several of the most common reasons for reprocessing are:

- to include data manually collected on a thumb drive from the sites (usually 1-2 weeks after data was originally collected)
- to apply updated and adjusted calibration values and or limits
- to account for missing input data (primarily for VAPs)
- as directed through DQRs, DQPRs, and EWOs.

4.3 Stored Phase

Data transfer from the DMF to the ARM Data Archive occurs nightly. All data, from raw to the highest processed level, are sent to the ARM Data Archive. The data transfer software that runs at the DMF ensures only data that pass numerous predetermined data checks are allowed to be delivered to the ARM Data Archive. Those that do not pass all data checks are flagged for further follow up.

Those data checks include: unexpected file splits, empty or zero length files, age of data, previously held files, correct tar file properties, correct NetCDF file properties, duplicate file storage, missing fields without QC flags, duplicate time samples with same NetCDF file, not a number (nan) values, fill values, infinite values, and values that are equal to or great than +10e10, and values that are equal to or smaller than -10e10 in size.

4.4 Available Phase

The ARM Data Archive receives data from other sources including the External Data Center (XDC) and the Reprocessing Center. The DMF manages data flow rates with the ARM Data Archive so that bandwidth and disk storage are not overburdened.

The Data Delivery Tracking (DDT) tool reports whether instruments have completed the Online, Processed, Stored, and Available phases related to data flow. The Available phase also states whether data are available at the ARM Data Archive complete with web and metadata documentation.

For additional information related to web documentation, metadata documentation, and ARM Data Archive services, please visit www.archive.arm.gov.

5.0 AMF Start-up and Support Associated with the Commission of New Sites.

The DMF plays a critical role in the product initialization and tracking of data flow from research site instruments and systems. For each deployment of an AMF or the initialization of a new research site, DMF staff will manage the associated tasks to meet operational milestones. This requires a thoughtful ramping up of tasks prior to operational start milestones. These efforts are governed and communicated within the Engineering Change Process. Tracking and communication details are carried out through DMF services.

6.0 General Operations

The following tasks will be done by the Operations staff:

- critical/emergency DMF support availability
- delegation and contact plan
- track data flow problems from their source through DQPR, EWO, ECR and BCR systems
- submit DQPRs for identified data concerns
- notify site ops of problems discovered
- notify data clients of data flow problems
- monitor data flow and both ingest and VAP processing routinely (during normal business hours)
  - using DSView processing interface at DMF
  - using the DDT tool

  - inspecting processing logs
  - inspecting site_transfer logs
  - inspecting clean, bundle, and other logs
  - monitoring automatic emails for problems
  - monitoring system loads
- pre-arranged weekend and after hours support available as needed
- coordinate minor near-real-time reprocessing efforts with RAB
- coordinate with system administrators regarding system functionality to mitigate events that could result in loss of productivity, loss of data, or degraded data quality Monitor/coordinate BCR/ECR implementation impacts
- track data disk shipments
- coordinate appropriate data retention plan for efficient processing of VAPs
- manage a budget and staffing to perform operations
- maintain data system configuration files for:
  - data transfers
  - data processing
  - DSView
  - calibration files
- work closely with VAP developers in release and support of VAPs
- work closely with engineering in release and support of production software
- release software according to ARM CM standards and BCR process
- perform as needed testing of software as part of the CM process
- monitor NSA weekly reports and daily SDS communication
- maintain inter-site software comparison and standardization tool.

The following reports will be generated by Operations:

- Quarterly DMF Report.

7.0 Summary

The DMF Operations staff are responsible for the processing and management of data from all sites to the ARM Data Archive. Operating the DMF not only includes working with staff located at Pacific Northwest National Laboratory (PNNL), but requires a close relationship with numerous ARM groups and affiliates external to PNNL. As a centralized facility with experienced data operators, the DMF is ideally suited to monitor, manage, and mitigate many data transfer and processing situations.

Appendix A
DMF Contacts/Delegates/Emergency Plan

A.1 DMF Primary Contacts

Nicole Keck
Data Management Facility Manager
Pacific Northwest National Laboratory
P.O. Box 999, MSIN K7-28
Richland, WA 99352
Nicole.Keck@pnnl.gov
Phone: (509) 375-6470
Cell: (509) 947-8298

Tonya Martin
Data Management Facility Technical Support
Pacific Northwest National Laboratory
P.O. Box 999, MS K7-28
Richland, WA 99352
Tonya.Martin@pnnl.gov
Phone: (509) 375-4384

Anthony Clodfelter
Data Management Facility System Administrator
Oak Ridge National Laboratory
1 Bethel Valley Road
Building 2040 RM E136 MS 6290
Oak Ridge, TN 37831
clodfelteraj@ornl.gov
Phone: (865) 576-6199

A.2 Alternative Delegates

Matt Macduff
Data Integration Group Technical Lead
Pacific Northwest National Laboratory
P.O. Box 999, MS K7-28
Richland, WA 99352
Matt.Macduff@pnnl.gov
Phone: (509) 372-4704

A.3 Emergency Contact Strategy

In the event of an emergency outside of normal business hours, all concerns should be directed to the DMF Manager, Nicole Keck, by email or cell phone. She will then contact the system administrator, technical support, and other ARM management as needed.

Appendix B
ARM Data Information Flow Diagram

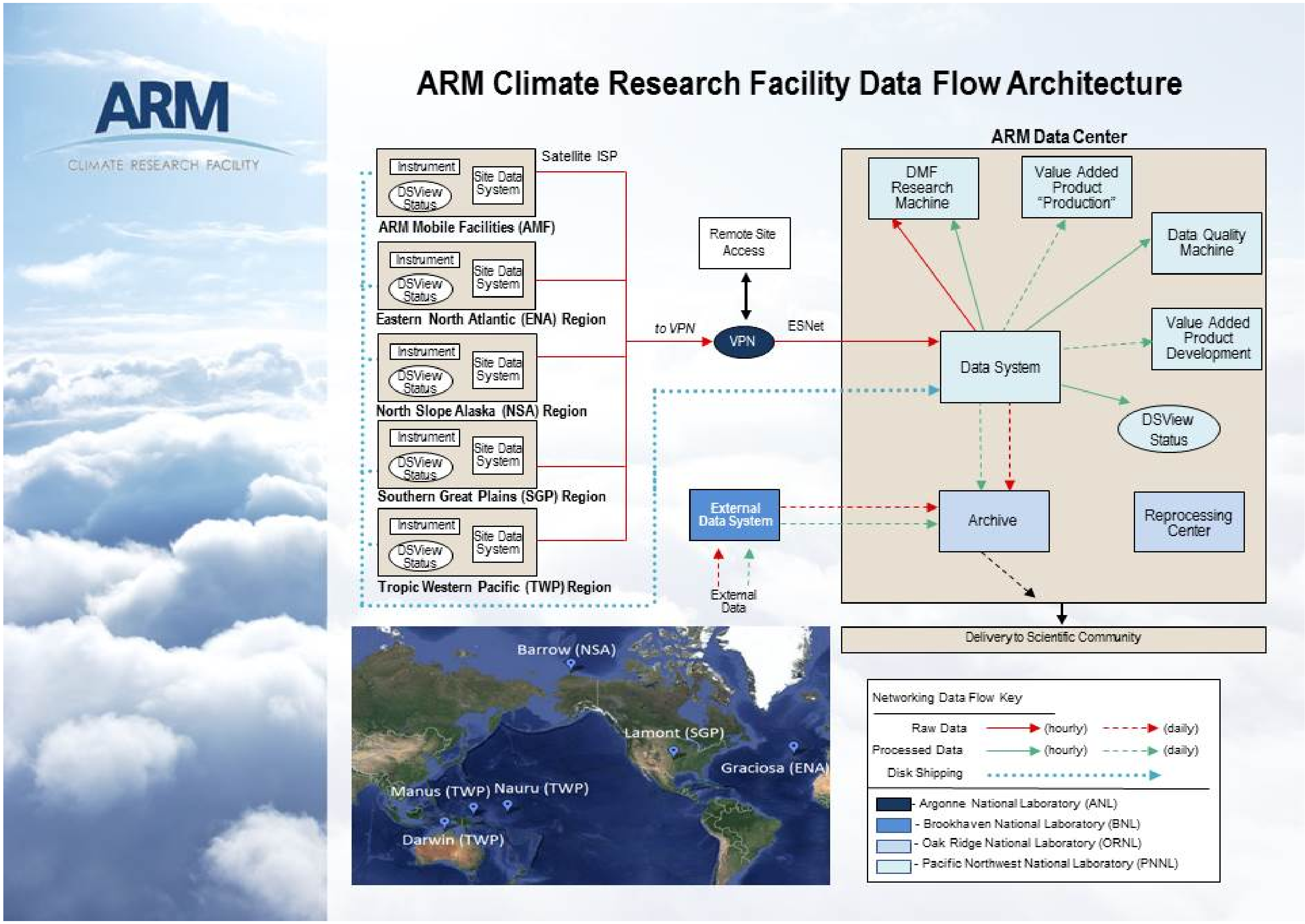
ARM
CLIMATE RESEARCH FACILITY
ARM Climate Research Facility Data Flow Architecture
ARM Data Center
Satellite ISP
Instrument
DSView Status
Site Data System
ARM Mobile Facilities (AMF)
Eastern North Atlantic (ENA) Region
North Slope Alaska (NSA) Region
Southern Great Plains (SGP) Region
Tropic Western Pacific (TWP) Region
Remote Site Access
to VPN
VPN
ESNet
DMF Research Machine
Value Added Product "Production"
Data Quality Machine
Value Added Product Development
Data System
DSView Status
External Data System
External Data
Archive
Reprocessing Center
Delivery to Scientific Community
Barrow (NSA)
Lamont (SGP)
Graciosa (ENA)
Manus (TWP)
Nauru (TWP)
Darwin (TWP)
Networking Data Flow Key
Raw Data (hourly) (daily)
Processed Data (hourly) (daily)
Disk Shipping
- Argonne National Laboratory (ANL)
- Brookhaven National Laboratory (BNL)
- Oak Ridge National Laboratory (ORNL)
- Pacific Northwest National Laboratory (PNNL)

Figure 2. ARM Data Flow Architecture

ARM

CLIMATE RESEARCH FACILITY

www.arm.gov

U.S. DEPARTMENT OF ENERGY

Office of Science