Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	Atmospheric Radiation Measurement (ARM) Data Center
Website:	https://www.arm.gov/data
Certification Date:	15 May 2020

This repository is owned by: **Oak Ridge National Laboratory**

Atmospheric Radiation Measurement (ARM) Data Center

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, National repository system; including governmental

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of Repository

The US Department of Energy's (DOE) Atmospheric Radiation Measurement (ARM) Data Center is a long-term archive and distribution facility for various ground-based, aerial, and model data products in support of atmospheric and climate research. The ARM facility currently operates more than 400 instruments at various observatories (https://www.arm.gov/capabilities/
observatories).

Working together, nine DOE national laboratories manage and operate the ARM user facility. This partnership supports DOE efforts to better understand and predict Earth's atmosphere in order to develop sustainable solutions to the nation's energy and environmental challenges. The collaborating labs are listed at
https://www.arm.gov/about/management-structure/labs.

The ARM Data Center (ADC) archive currently holds more than 11,000 data products with a total holding of over 1.9 petabytes of data that dates back to 1992. These include data from instruments, value-added products, model output, field campaigns, and data contributed by principal investigators (PIs).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

Atmospheric scientists around the world use ARM data to conduct research to advance process-level understanding of the key interactions among aerosols, clouds, precipitation, radiation, dynamics, and thermodynamics, with the ultimate goal of reducing the uncertainty in global and regional climate simulations and projections.

The ADC currently archives data from three long-term atmospheric observatories that include sites located on the Southern Great Plains (Oklahoma), on the North Slope of Alaska (Barrow–Utqia■vik), and in the eastern North Atlantic [the Azores]). To explore research questions beyond those addressed by ARM's fixed atmospheric observatories, scientists can propose a field campaign to use one of three ARM mobile facilities to collect atmospheric and climate data from under-sampled regions around the world. The ADC currently archives data from all these facilities and distributes it to atmospheric scientists around the world.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

A. Content distributed as deposited, B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation, D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

ARM measurements are aggregated in data files. A data file usually contains a time series of one or more measurements for a known time interval and a single location. Most data are stored in the NetCDF format.

Data levels are based on the "level of processing," with the lowest level of data being designated as raw or "00" data. Each subsequent data level has minimum requirements; and a data level is not increased until ALL the requirements of that level, as well as the requirements of all data levels below that level, have been met.

00
raw data—primary raw data stream collected directly from instrument

01

raw data—redundant data stream or sneakernet data

a0

converted to NetCDF

a1

calibration factors applied and converted to geophysical units

a2… to a9

further processing on a1-level data that does not merit b1 classification

b1

QC checks applied to measurements

b2… to b9

further processing on b1-level data that does not merit c1 classification

c0

intermediate value-added data product; this data level is always used as input to a higher level “VAP”

c1

derived or calculated value-added data product (VAP) using one or more measured or modeled data (a0 to c1) as input

c2… to c9

further processing applied to a “c1” level data stream

s1

summary file consisting of a subset of the parent .c1 file with simplified QC and known ‘bad’ values set to missing

s2

summary file consisting of further-processed s1 data.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

Not Applicable

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

Further details are available at:

ARM web site: https://www.arm.gov/

ARM Data Center: https://www.arm.gov/data

ARM Monograph: https://journals.ametsoc.org/toc/amsm/57

ARM Decadal Vision: https://www.arm.gov/publications/programdocs/doe-sc-arm-14-029.pdf

ARM Data Citation Paper: https://www.mdpi.com/2306-5729/1/2/11/htm

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

As part of the ARM User Facility, ADC's mission is to provide processing, submission, archival, and distribution of ARM data to the scientific community and the other archives. These data are generally available within 48 hours of collection. The total archive contains over 1.9 petabytes of data from over 11,000 data products. ADC delivers about 30 terabytes of data per month through Data Discovery.

Additional Resources:

ADC mission/scope:
• ARM mission: https://arm.gov/about/mission-and-vision-statements
• https://www.arm.gov/data/work-with-arm-data
• https://www.arm.gov/connect-with-arm/organization/arm-data-center
• ARM Data Policy: https://www.arm.gov/policies/datapolicies/

Link to the DOE approval document for ARM and ADC
- https://science.osti.gov/ber/Facilities/User-Facilities/ARM

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

ARM data are freely available to any user in the world. However, users need to register with ARM to download the data. This information is required for documenting the value and usage of the ARM facility by the research community. As a DOE scientific user facility, ARM welcomes users from all institutions and nations.

As stated in the ARM Data Policy (https://www.arm.gov/policies/datapolicies/datasharingpolicy), all data obtained through ARM are monitored for quality and made available free of charge through the ADC. However, users are required to register (free of charge) before downloading the data; this helps ARM to communicate any information regarding quality assessment or reprocessed data availability, as well as gather usability metrics to satisfy national user facility requirements. There are no restrictions on the use of ARM-collected data published in any publicly available resources, which implies there is no restriction. The ADC recently applied Creative Commons and is in the process of updating the data policy to include the statement “The ARM data will be released publicly under a Creative Commons Attribution 4.0 license (CC-BY) license.”

Anytime a user requests ARM data, an order notification email is sent to the user explaining how to download data using various freely available protocols (FTP, THREDDS, web service, and so on). In the notification email, users are "encouraged" to cite the use of ARM data and acknowledge the ARM program. The data citation guidance (https://www.arm.gov/working-with-arm/acknowledging-arm/doi-guidance-for-datastreams) is encouraged as a best practice as part of data management and not a requirement; therefore, no mechanism is employed to enforce the use of citation. No ARM publicly available resources explicitly say that citing data is a requirement.

Beginning in fiscal year 2015, the DOE Office of Science (SC), the primary sponsor of ARM, has required that a limited set of information relating to a user project/experiment be transmitted to DOE SC. At the conclusion of each fiscal year, a subset of this information—including user name, institution affiliation(s)and project title(s)—is publicly disseminated as part of an SC user projects and experiments database on the SC website (http://science.energy.gov). For proprietary projects, DOE SC requests that the user provide a project title suitable for public dissemination.

Useful link: ARM Data Policy: https://www.arm.gov/policies/datapolicies/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All the ARM facility–collected and PI-contributed data are available via the ADC. A secondary copy of all data is deep archived using Oak Ridge National Laboratory (ORNL) mass storage system. In addition, another copy of all ARM data is archived at the Argonne National Laboratory (Argonne) High Performance Storage System (HPSS).

The ADC (previously known as the ARM Archive) has been operational since 1993 (https://journals.ametsoc.org/doi/full/10.1175/AMSMONOGRAPHS-D-15-0043.1). In its 26 years of continuous operation, ADC has implemented many elements of "backup and preservation" in its operation and design. These include the following:

1. Backup of ARM data files

2. Backup of field campaign and PI data files
3. Backup of ARM Archive databases
4. Backup of ARM systems and Archive software

The ARM’s designation as a DOE national user facility ensures the long-term preservation of all collected data. In case of a major change such as facility closure, the potential caretaker of the data for future preservation will be DOE's Office of Scientific and Technical Information (OSTI). ARM has already established sharing of metadata and technical documents with OSTI. In case of major changes within the institution, the data also are available from Argonne. Any further preservation (in case of ADC closure) will be implemented based on guidance by the funding agency (DOE).

Policies and procedures for data archival and release for the ARM Facility are pursuant to those of the US Global Change Research Program (USGCRP), as described on the USGCRP website: http://www.globalchange.gov. Specific details can be found in the USE DOE document (DOE/SC-ARM-13-022) for ARM Facility Management available at https://www.osti.gov/servlets/purl/1253897

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

0 – Not applicable

Reviewer Entry

Reviewer 1

Comments:
0 – Not applicable

Reviewer 2

Comments:
0 – Not applicable

Response:

ARM datasets contain no personal information and hence present no disclosure risk. The privacy and security notice is presented at https://www.arm.gov/policies/privacy-security-notice

ARM-sponsored data are released in the general spirit of the basic tenets of the ARM user facility:

1. Free and open access.
2. Immediate processing and sharing by PIs in the field, if at all possible.
3. Timely release to collaborating science teams and the general scientific community through the ARM data system.
4. All data to be submitted to the ARM data system will be accompanied by full documentation in accordance with the Data Management and Documentation Plan. Further details are available at :
https://www.arm.gov/policies/datapolicies/generalguidelines

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The ADC (previously the ARM Archive) is a long-term archive (since 1993) operated by DOE's ORNL. The ADC is funded by the DOE Biological and Environmental Research (BER) program. About 22 staff members support ADC operations. The program is sufficiently funded to carry out the mission. The ADC also gets additional funds from DOE BER to maintain and upgrade computing and data storage. The funding provided by DOE is project-based and operational support for the ARM program, which is one of DOE's national user facilities.

The ADC has a robust group of staff with over 200 years of cumulative experience in end-to-end science data management. The hosting institution, ORNL, recently celebrated 75 years of great science (https://www.ornl.gov/content/seventy-five-years-great-science).

Within ORNL, the ADC is part of the Environmental Sciences Division. Many ADC staff work in multiple data repositories that are collocated with ARM Data Center. ADC staff are listed at https://www.ornl.gov/group/arm-data-science-and-integration/staff.

The ADC staff are all located in the same facility and meet regularly to resolve various issues and brainstorm new ideas. They also regularly participate in various data and computing management training events and workshops (a minimum of two a year). Some examples are cybersecurity training, software training (e.g., Python, NoSQL Java), database trainings leadership training, SciPy conferences, big-data workshops (Devnexus and ApacheCon), and various ORNL-hosted in-house training types.

ADC staff are regular and active participants in various international data and computing bodies, including ESIP, GEO, DataOne, AGU, and CODATA. Two ADC members are also IEEE (Big Data) senior members.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

ARM conducts a triennial review using external subject matter experts identified by DOE. In addition, the ADC receives recommendations from the ARM User Executive Committee (UEC, https://www.arm.gov/about/constituent-groups/uec). The UEC provides objective, timely feedback to ARM leadership with respect to the user experience. The committee serves as the official voice of the user community in its interactions with ARM management. Additional recommendations and guidance are provided by the following groups:

• The ARM Infrastructure Management Board (IMB; https://www.arm.gov/connect-with-arm/organization/infrastructure-management-board-imb) provides guidance on managing ARM with input from the research community and ARM staff.

• The ARM-Atmospheric System Research (ASR) Coordination Team (AACT; https://www.arm.gov/connect-with-arm/organization/aact) is a constituent group to foster communication between ARM leadership and users, ASR scientists, and DOE program managers.

• The Architecture and Services Strategy Team (ASST; https://www.arm.gov/connect-with-arm/organization/asst) is responsible for the representation of and communication with the software development and operations team members. ADC staff also interacts with various ASR working groups during the annual science team meetings and gathers various feedback and recommendations.

Reviewer Entry

Reviewer 1

Comments:

Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

ARM Data Center follows end-to-end data security and integrity. After the data collected from instruments that are deployed in remote locations, the data gets sent to ORNL via a secured and encrypted transfer method. When data arrives to ORNL, all incoming data gets scanned. Proper md5 checks are carried out during the data processing pipeline which ensures data integrity.

Different versions of data are deep archived (raw data from instruments, processed data, value added products etc.) at the ORNL Mass data storage system (HPSS), a LTO tape based copy also gets created a back up, an additional offsite data back up is done using the Argonne National Laboratory's HPSS.

Data Provence, versioning details, and change histories are stored in various metadata databases. A dedicated team handles the metadata workflow and ensuring the completeness of metadata.

Robust automated processes are used in continuously monitoring the data and metadata, these automated workflows asses completeness of data (using automated MDF5 checks before and after transferring) and metadata (using the checks applied in the metadata management tool and standards compliance engine). Please refer the attached diagram

that has high level detail of this.

All the data collected using the instruments go through a detailed semi-automated quality analysis before users can access the data. If a data quality issue is identified, a Data Quality Report (DQR) gets created by the reviewer (instrument specialist, data developer or the ARM data quality office). Data Quality report are captured in the database.

Below are the data versions of typical ARM collected data files:

00: raw data – primary raw data stream collected directly from instrument

01: raw data – redundant data stream or sneakernet data

a0: converted to netCDF

a1: calibration factors applied and converted to geophysical units

a2… to a9: further processing on a1 level data that does not merit b1 classification

b1:QC checks applied to measurements

b2… to b9: further processing on b1 level data that does not merit c1 classification

c0: intermediate value-added data product; this data level is always used as input to a higher level “VAP”

c1: derived or calculated value-added data product (VAP) using one or more measured or modeled data (a0 to c1) as input

c2… to c9: further processing applied to a “c1” level data stream

s1: summary file consisting of a subset of the parent .c1 file with simplified QC and known ‘bad’ values set to missing

s2: summary file consisting of a further – processed s1 data.

ARM uses NetCDF data convention, where feasible, it uses Climate Forecasting (CF) data standard. ARM data are distributed using variety of protocols including FTP, THREDDS/OpenDAP, DropBox, GlobusOnline etc.. ARM Data Center distributes metadata to other portals and learning houses using ISO-19115, FGDC standards, Schema.org, and JASON-LD.

When data gets reprocessed, the changes are captured in the Data quality report, users who ordered these data in the past gets notified about the data change (due to the reprocessing). The producers (instrument mentors and data

developers) are involved in the reviewing the changes and confirming the results and documenting the changes.

The ARM Data Center maintains the data provenance using various databases. The provenance information is constantly used in reprocessing and new data ingests. Some of the provenance include: input parameters, source data streams, software versions used, algorithms used, version details, data quality checks applied etc.. The provenance details that are deemed useful to the users are embedded in the data file as a global attribute (for both human and machine readable).

The ARM Data Center maintains and provides additional metadata via the ARM data discovery tool, for example: for a given data stream, users can click a link provided to view the instrument handbooks and data technical notebooks. These documents are maintained by the ARM facility suing the instrument specialist and science data translators.

DOIs are applied to datasets and a citation generator service is provided within the data discovery (please refer:https://www.mdpi.com/2306-5729/1/2/11/htm)

Further details can be viewed in the ARM monograph , published in the AMS (https://journals.ametsoc.org/doi/full/10.1175/AMSMONOGRAPHS-D-15-0043.1)

Following is an example of data provenance shipped in every datafile:

// global attributes:
:process_version = "vap-qcrad1long-6.1-0.el6" ;
:dod_version = "qcrad1long-c2-2.0" ;
:command_line = "qcrad1long -n qcrad1long_sirsc2 -s sgp -f C1 -d 20180321 -D -R" ;
:site_id = "sgp" ;
:facility_id = "C1: Lamont, Oklahoma" ;
:datastream = "sgpqcrad1longC1.c2" ;
:qc_standards_version = "1.0" ;
:Title = "Data Quality Assessment of ARM Radiation DATA" ;
:BEGSW = "BestEstimate_down_short_hemisp" ;
:MFRSRGSW = "MFRSR_hemisp_broadband" ;
:UC1 = "First user configurable limits" ;
:UC2 = "Second user configurable (extremely rare) limits" ;
:PP = "BSRN physically possible limits" ;
:DirN = "short_direct_normal" ;
:DifSW = "down_short_diffuse_hemisp" ;
:SumSW = "DirN * cos(SZA)+ DifSW" ;
:GSW = "down_short_hemisp" ;
:LWdn = "down_long_hemisp" ;
:LWup = "up_long_hemisp" ;
:SWup = "up_short_hemisp *Alert: early Nauru (twpC2) data was over a bright non-representive surface, data are set to

-9999 as a result" ;
:sigma = "Stephan-Boltzmann constant = 5.67 * 10 ^ -8" ;
:SZA = "Solar Zenith Angle" ;
:Ta = "Air temperature" ;
:Tmin = "User defined minimum air temperature" ;
:Tmax = "User defined maximum air temperature" ;
:Tsnow = "Temperature limit for albedo limit test, temperature at which snow limit is allowed" ;
:Tracker_off_limit = "0.850000" ;
:Tc-Te_bimodel = "Tc and Te differences for separating wet and dry modes = 6.0" ;
:postprocessing_description = "The c1 level datastream uses generic average correction coefficients (based on analysis of many years of actual site pairings) for IR loss in the unshaded pyranometer measurements. The c2 level datastream uses more advanced correction coefficients calculated by the gswcorr1dutt VAP. Radiometers are replaced each year and c2 level correction coefficients are calculated for each radiometer pair (PSP and collocated PIR) separately for each site for each deployment year. Variables potentially affected by the updated correction coefficients are: down_short_hemisp and BestEstimate_down_short_hemisp and their associated qc and aqc flags" ;
:Dutton_correction_dry_coefficient = "0.033528" ;
:Dutton_correction_wet_coefficient = "0.051985" ;
:Full_correction_dry_coefficient_b1 = "0.019724" ;
:Full_correction_dry_coefficient_b2 = "1.378032" ;
:Full_correction_wet_coefficient_b1 = "0.032314" ;
:Full_correction_wet_coefficient_b2 = "1.318913" ;
:configurable_limits = "\n",
"cnf0 = 9.0 * snow covered ground Ta limit for albedo tests (real, degrees C > 0.0)\n",
"cnf1 = 0.92 0.97 * Max GSW climatological mult. limit factor (real < 1.2)\n",
"cnf2 = 0.52 0.58 * Max DifSW climatological mult. limit factor (real < 0.75)\n",
"cnf3 = 0.82 0.86 * Max DirNSW climatological mult. limit factor (real < 0.95)\n",
"cnf4 = 0.87 0.95 * Max SWup climatological albedo limit factor (real < 1.0)\n",
"cnf5 = 190.0 145.0 * Min LWdn climatological limit factor (real > 60.0)\n",
"cnf6 = 465.0 500.0 * Max LWdn climatological limit factor (real < 500.0)\n",
"cnf7 = 240.0 210.0 * Min LWup climatological limit factor (real > 60.0)\n",
"cnf8 = 590.0 630.0 * Max LWup climatological limit factor (real < 700.0)\n",

:input_datastreams_description = "A string consisting of the datastream(s), datastream version(s), and datastream date (range)." ;
:input_datastreams_num = 32 ;
:input_datastreams = "sgpgswcorr1duttC1.c1 : 1.8 : 20180314.000000-20180321.000000\n",
"sgpmetE13.b1 : 4.37 : 20180314.000000-20180321.000000\n",
"sgpmfrsrC1.b1 : 11.11 : 20180314.000000-20180321.000000\n",
"sgpsirsC1.b1 : 12.0 : 20180314.000000-20180321.000000" ;
:doi = "10.5439/1227214" ;

:history = "created by user howie on machine amber at 30-Jan-2019,06:39:13" ;

}

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

ARM uses a robust data collection structure to identify community needs for various atmospheric parameters using science working groups. Then a instrument plan is developed along with the data collection policy. Full details are available at the arm.gov web page or in the monograph at https://journals.ametsoc.org/toc/amsm/57.

The procedure for PIs to submit ARM science research products, field campaign data, or DOE-supported research data to the ARM Data Center is listed below:

• First, use the Data Product Registration and Submission form.
• Select a data type. This choice determines the level of review and the procedure for handling and approving the

documentation and accompanying data submissions within the ARM user facility. Currently, three data type options are available.

1. ARM field campaign data—Reviewed by ADC staff responsible for handling field campaign data submissions.
2. ARM PI data product—Reviewed by ARM translators and infrastructure representatives.
3. Research data for the ARM—Sent directly to metadata reviewers for the ARM Archive.

The Data Product Registration and Submission form uses the FGDC and provides various functionalities to PIs to properly describe their data using commonly available keywords (e.g., CF, ARM primary measurement names).

If the metadata provided is insufficient, the metadata experts team communicate with the data submitter to complete the missing details.

The review of data quality is centrally managed by the ARM data quality office, which uses a web-based ticket system to capture, resolve, and report data quality issues. This system is managed by the data quality office, but problems are solved collaboratively with instrument technical staff and site operations.

All the data collected using the instruments go through a detailed semi-automated quality analysis before users can access the data. If a data quality issue is identified, a DQR its created by the reviewer (instrument specialist, data developer, or ARM data quality office).

More about the ARM Data Quality Program can be found at https://www.arm.gov/data/data-quality-program and http://cimms.ou.edu/wp-content/uploads/2018/11/ARM-Program-Data-Quality-Office.pdf

For data products submitted by PIs, the Data Product Registration and Submission form requests the inputter to add data quality details. A PI who submits non-preferred (non-NetCDF) format data will have to describe how to use the data and what tools are available to visualize the data.

Once the PI submits the metadata and data using the Data Product Registration and Submission form, the system checks for completeness. Then the system automatically starts the review process, which includes review by the ARM metadata experts, science liaison, and communications group. All metadata submitted by the PIs go through a rigorous vetting process by the ARM metadata team. For data-insufficient metadata, the team assign additional metadata needed for long-term preservation. If needed, the review team returns the submitted metadata to the PI to add details. When approved, these data are properly archived, DOIs are assigned (where applicable), and they are made available via the data discovery tool.

ARM-collected data are made available in NetCDF formats, and PIs are encouraged to submit their data using NetCDF as well.

ARM data formats are published at https://www.arm.gov/data/work-with-arm-data.

When NetCDF format is not possible, PIs can submit data in their native format. They are expected to provide additional details about how to use the data and any tools available to read the data, and so on, using the Data Product and Registration Tool.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The ADC archives all raw, processed, value added data products using data versions. The current archival and data backup process employed by the ADC consists of multi-tier data storage. There are two production copies of all ARM data in the ORNL mass storage system (HPSS). ORNL HPSS currently uses a big online disk cache and LTO-based tape backups. In addition, all incoming data are stored in LTO tapes and stored in a separate building within ORNL. Beginning in 2018, the ADC also started creating a true offsite backup by sending all the incoming data to the Argonne HPSS located in Chicago. Previously, the ADC had created a full copy of all historical data on LTO tapes and sent them to Argonne. The data transfer is done through a secured ESNET using the Globus transfer protocol.

In addition to those archived copies, the ADC keeps a large volume of current and popular data streams in an online data storage system for immediate access. The current online storage capacity is about 2 petabytes, and over 900 terabytes of data are currently stored for immediate access.

Automated processes check for data integrity and consistency across the archival copies using the md5 checksums. If data are corrupted in one storage copy, the system fetches from the next available archived copy (from ORNL HPSS or Argonne HPSS copies).

To handle media deterioration, LTO tapes used to store the data (either ADC tape backup or HPSS storage) are routinely upgraded before the end of their life. For example, ORNL HPSS, as standard operational procedure, upgrades its tapes by automatically moving the data from tapes to a built-in large disk cache (~20 petabytes). Thus the tapes can be upgraded without any outage. (More detail about the about ORNL HPSS is at https://www.olcf.ornl.gov/olcf-resources/data-visualization-resources/hpss/.)

All the data archival and backup processes are well documented and used for operational purposes.

ARM has developed a highly integrated operating system to support measurements at its ground-based and aerial facilities. Data systems shared across the ground sites have the same configurations; and all data are processed at the same data management facility, which is now co-located with the data archive from which data are made available to the user community.

Just as the ARM ground sites share many of the same instruments, each site also uses the same design for its site data system. Each site has a data system that serves as the central nervous system for the facility. All data are collected by the site data system and the data system is used to manage the transfer of data from the sites to the ADC at ORNL. These data systems have been designed to be interchangeable across sites, differing only in the amount of storage available, with the most remote sites allocated greater storage. Sharing site data systems design and resources prevents each of the site operations laboratories (Argonne, Los Alamos, and Sandia) from needing local site data system personnel to manage the operation of these critical systems. Proper md5 checks are carried out throughout the entire data pipeline to ensure data integrity and consistency when data are transferred from one process to another.

Highly experienced staff manage the data services operations, which include the site data systems, the ADC, and the data quality office. ARM has developed and implemented a variety of management processes and associated tools to ensure that instruments are operated and that data are collected and processed so as to provide high-quality data to the user community. Three areas that have a particularly direct impact on the ARM user community are change management, data quality, and the collection and processing of ARM data.

Further details can be accessed from the ARM Decadal Vision: https://www.arm.gov/publications/programdocs/doe-sc-arm-14-029.pdf

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The data collected as part of the ARM mission have been preserved for the last 25+ years, with the first dataset dating back to 1992. A highly structured preservation process is followed to make sure data will be preserved for the long term, using archival storage of the data explained in Sections IX — R9. All data including (1) raw, (2) calibrated, (3) ingested, (4) value added, (5) synthesis, (6) field campaign, and (7) PI-submitted data are preserved and distributed using a well-matured workflow. The ADC is the official data repository for the data collected by ARM. PI-submitted data, with the approval of the PIs, can be made available to users with the appropriate DOIs, along with the recommended data citation. The ADC has the rights to copy, transform, and store the archived ARM data.

ADC staff ingest the data, prepare metadata for discovery, and compile comprehensive documentation in the form of technical reports and instrument handbooks for future users; we use the commonly used 20-year data preservation rule—a time far enough into the future to be useful for preparing the documentation for both sharing and archiving data. Data are preserved for the future following DOE’s data preservation guideline.

The funding agency (DOE) website (https://science.osti.gov/ber/Facilities/User-Facilities/ARM) mentions the ADC (previously known as ARM Archive). Also, the ARM website (https://www.arm.gov/about/stats/reporting-requirements) states “The ARM Data Center is the repository of all ARM data and all user facility data.“ In addition, a detailed monograph (https://journals.ametsoc.org/toc/amsm/57) explains ADC best practices for the first 20 years of the ARM program; specifically Chapter 11, ARM Data System and Archive (https://journals.ametsoc.org/doi/full/10.1175/AMSMONOGRAPHS-D-15-0043.1), contains details regarding how data services evolved during the first 20 years of the program.

The DOE ARM Climate Research Facility Management Plan (DOE/SC-ARM-13-022; https://www.osti.gov/servlets/purl/1253897) states in Section 9 that preservation of ARM data is pursuant to USGCRP guidance.

The ADC home page on the ARM website refers to the functional responsibility of ADC: https://www.arm.gov/connect-with-arm/organization/arm-data-center

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

A comprehensive data quality assessment program is essential for documenting the quality assurance process and ultimately in producing a dataset of some prescribed known quality and usability. The program must collect and track data about the system (metadata) at every point along the assessment path, from instrument selection and procurement to initial fielding and beta testing; to field operation, data collection, and quality inspection and assessment; to problem reporting and resolution; and to data distribution and communication of information about those data. If the process and details of the data created cannot be described, the data will have limited scientific value.

ARM follows a well structured and documented data and metadata quality process. The details are provided in the following paragraphs.

Given the data volume of over 1.9 petabytes (as of April 2020), data inspection and assessment activities must be automated and efficient, although human inspection of the results still remains a high priority.

The quality assurance model has three components. The first component is a “rapid evaluation and response” piece involving data inspection and assessment. It is designed to identify gross and some more subtle issues within the data streams as fast as possible and relay that information to site operators and the instrument mentors so that the (potential) problem-resolution process can begin. The goal of this component is to minimize the amount of data that is affected by the problem. The second component involves documenting and reporting data quality issues for the scientific user; this is primarily done via text-based but machine-readable DQRs (item 3 in the list below). The third component involves reprocessing of data after known problems have been identified and solved to provide end users with the best products available.

Once data have been inspected and assessed, a variety of reporting mechanisms allow the data quality analysts to inform instrument mentors, site operators, and site scientists of their findings. Data quality reporting mechanisms are based on web-searchable and -accessible databases that allow the various pieces of information produced during the quality assurance process to be neatly conveyed to problem solvers in a timely manner. The system described below results in even and consistent treatment of ARM problem reporting and resolution.

The problem reporting system is divided into three linked processes:

1. Weekly reports are issued on data inspection and assessment by analysts from the data quality office and distributed internally to instrument mentors, site operators, and site scientists.

2. Reports are issued describing problems discovered by data quality analysts or instrument mentors and distributed internally to instrument mentors, site operators, and site scientists so they can initiate a problem-resolution process. These online reports document the progress and status of the actions proposed and implemented.

3. DQRs documenting a known problem and its resolution, written by instrument mentors, are distributed publicly to the data user community.

The ARM infrastructure conducts an extensive data reprocessing program that is informed by the data quality assessment process. Reprocessing is performed to fix known data issues and has been used extensively throughout the lifetime of ARM. Reprocessing requires the modification or elimination of previous DQRs and the subsequent reissuing of data to all who may have downloaded the data from the data archive.

All the generated data quality information is provided to users in different phases of data discovery, downloading, and use. Users who ordered the data in the past will be notified if data quality information is available for the data they received in the past. Users can provide feedback and input about the data quality using the online feedback system by selecting "Data Quality issues" in the feedback. The system automatically sends the feedback to the ARM data quality experts.

References to external data products and citations to other relevant data products are captured in the ARM publication database (https://www.arm.gov/research/publications).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

ARM has an end-to-end data operations system that enables the flow of data from the observation facilities to the scientific community. This system comprises core components such as the site data system, the data flow software, and the ADC. The ADC workflow includes data ingestion and processing, data reprocessing, metadata workflow and field campaign data management, data archival and backup, data discovery and distribution, database and workflow, and the data archive. The flow through this system is complex, and several software tools have been developed to track the status of data, particularly new data streams, as they work their way through the system. While these tools provide very useful automated reports, ultimately, problem solving requires close communication among the data groups; and management processes have been constructed to maximize the efficiency of this complex system.

Each observation site has a site data system that serves as the central nervous system for the facility. All data are collected by the site data system, which is also used to manage the transfer of data from the sites to the ADC at ORNL. These data systems have been designed to be interchangeable across sites, differing only in the amount of storage available, with the most remote sites allocated greater storage. It is required that each site have at least 6 months of storage to protect against the possibility that the site could become temporarily cut off from off-site data transfer, or as contingency against a problem occurring with transfer, such as corruption of data. There may be other small differences as well; for example, for the recently concluded second mobile facility on a cruise across the Southern Ocean, the system was outfitted with solid state hard drives to mitigate the risk of data loss in that harsh environment. In addition, data, such as model and satellite data, are collected from external providers and ingested on the ADC data system. Once collected by the site data systems, data are next transferred to the ADC. For most data, this transfer is carried out automatically via the network. However, some of the largest datasets and some of the most remote sites require manual shipping of hard disks.

At the ADC, the files are checked for successful transfer and are processed to a standard format, which is an ARM standardized structure using NetCDF (http://www.unidata.ucar.edu/software/netcdf/). NetCDF is a self-describing binary format with many compatible software tools. Once processed, the data are catalogued, stored in the ARM archive, and made discoverable through association with an array of metadata characterizing information such as location and measurement classification. These metadata enable powerful searching capabilities through the data archive data discovery interface.

The process of characterizing and classifying metadata has been streamlined with the use of tools, workflows, and coordination calls. The principal tool for creating, assigning, and editing metadata is the Metadata Services Tool. It has multiple components that are used to assign existing metadata to standardized ARM baseline data, ARM Field Campaign (IOP) data, or ARM PI data. In addition to linking metadata to new data streams, this tool has functions to add new metadata to existing database tables. Processes for metadata creation and assignment are documented in the form of workflows.

Further details of the workflow are published at the following locations:

McCord, R. and J. Voyles, 2016: The ARM Data System and Archive. Meteorological Monographs, 57, 11.1–11.15, https://doi.org/10.1175/AMSMONOGRAPHS-D-15-0043.1

ARM Climate Research Facility Management Plan (DOE/SC-ARM-13-022), https://www.osti.gov/servlets/purl/1253897

The document at https://www.arm.gov/publications/programdocs/doe-sc-arm-14-029.pdf contains the details of data workflow, including various data services components such as data discovery, generation of data products, DOIs and citations, and security of data and software.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The ARM Data Center offers the very powerful Data Discovery tool (https://www.archive.arm.gov/discovery/). The discovery tool allows users to search and find over 11,000 data products using keywords and spatial and temporal filters.

Data Discovery was developed using modern and scalable architecture and is continuously upgraded to meet and exceed user expectations and stakeholder requirements. The discovery tools allow users to perform various data/metadata searches, view data quality information, and view data plots. The ADC recently enabled data access and delivery options such as THREDDS/OpenDAP, GlobusOnline, near–real-time data access API, automated data access via web services, advanced visualizations, a big data analysis platform for identifying data of interest, and the ADC's two high-performance computing clusters that allow scientists to access and conduct research using any archived ARM data. The ADC also recently upgraded the user registration process, tracking and reporting data usage including the download metrics and citations.

For external data portals, such as Google data search and data.gov, the ADC provides metadata in various protocols (e.g., ISO-19115, JASON, FGDC, OAI). The harvested metadata records are made available in these broader data clearinghouses. The ADC is currently in discussion with numerous broader data portals, such as Dataone and Polar data systems, to allow them to harvest ARM metadata records. Recently, the ADC provided an automated machine-harvestable metadata access point to Google data search.

The repository has been registered with https://www.re3data.org, data.gov etc..

Persistent identifiers:
The ARM data archive established a data citation strategy based on DOIs for ARM datasets to facilitate citing continuous and diverse ARM datasets in articles and other papers. This strategy eases the tracking of data provided as supplements to articles and papers. Additionally, it allows future data users and the ARM Climate Research Facility to easily locate the exact data used in various articles. Traditionally, DOIs are assigned to individual digital objects (a report or a data table), but for ARM datasets, a DOI is assigned to an ARM data product. This eliminates the need for creating DOIs for numerous components of the ARM data product, in turn making it easier for users to manage and cite ARM data with fewer DOIs. In addition, the ARM data infrastructure team, with input from scientific users, developed a citation format and an online data citation generation tool for continuous data streams. Further details are available at https://www.mdpi.com/2306-5729/1/2/11

The ADC offers recommended citations to the users. These are available at https://www.arm.gov/working-with-arm/acknowledging-arm/doi-guidance-for-datastreams

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

For data generated by ARM instruments, the metadata team creates a detailed metadata record along with technical documentation. These metadata records are made available in the data search and for distributing metadata. PIs who submit their data to ARM use the Online Metadata Editor (https://www.arm.gov/policies/datapolicies/data-product-registration-and-submission) to submit their metadata in FGDC and ISO format.

The ADC provides metadata to external/broader portals using the ISO 19115-2 and FGDC formats. In some cases, it also provides metadata in JASON and in custom formats needed by the external portals.

ARM data are primarily provided in the NetCDF format, as most of the climate research user community uses this format. In addition, users can request data in the ASCII format (using Data Discovery). For high-level products, ARM uses the CF convention.

The ADC provides a variety of additional resources to understand data, including structured DQRs (in human- and machine-readable formats), data plots, statistical summaries, big data analytical platforms, instrument handbooks, technical reports, and detailed metadata information. ADC also provides the key contacts (instrument mentors and data translators) in the Data Discovery and instrument pages.

Various measures are taken to account for the possible evolution of formats. Members of the ARM metadata team actively

participate in various metadata standards community discussions (for example, the NetCDF CF workshop at ESIP 2019). The ADC continuously adapts to standards evolution. For example, the Google data search tool and DataOne needed a specialized JASON (schema.org) format, and the aerosol user community wanted a specialized data export function (from NetCDF to ASCII format). These were successfully developed and offered by the ADC engineering team. All ARM metadata are captured in the database, which allows the ADC to modify and improve as metadata standards evolve over time.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The ADC has a state-of-the-art information technology infrastructure with a robust technical architecture.

For example, ARM uses streamlined and secured data transfer from various experimental facilities to the ADC, automated processing and archival processes to handle multiple terabytes of data per day, intuitive data discovery tools, a variety of data delivery options to handle small and large data transfers, advanced visualization, and a big data analysis platform for identifying data of interest.

In addition, the ADC has established two high-performance computing clusters (https://www.arm.gov/capabilities/computing-resources) using ORNL's leadership computing capability. These clusters allow scientists to access and conduct research using over 1.9 petabytes of archived data. The latest big data technologies, microservices, data citation strategies, and seamless development and analysis platforms were successfully applied to data center operation. Recently, the ADC deployed Jupyter notebook capability for users to develop and run their data analysis using the ARM high-performance computing clusters.

The ADC software stack includes a variety of open source-based processes, including Python and Java. A complete list of supported software within ADC research computing can be accessed at https://adc.arm.gov/tutorials/cluster/stratusclusterquickstart.html#available_software.

Two software repositories are maintained by the ARM data services, one for external community use (https://github.com/ARM-DOE) and one for infrastructure use (internal only).

For real-time datastream access, the ADC provides ARM Live Data Web Service: https://adc.arm.gov/armlive/. The network is backed with DOE's ESNET, which supports high bandwidth.

ARM tools and services follow community standards such as CF, FGDC, OpenDAP, REST, and SOAP.

Open-source and community-developed tools and standards are widely adopted at the ADC.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

[Note: Additional details about ARM Data Center's cybersecurity and incident response plan were shared with the CoreTrustSeal reviewers and Board, but have been removed owing to their sensitivity.]

A very strong security policy is in place for the successful operation of the data center. The ADC is within DOE's ORNL, which provides very strong security systems. In addition, the monitoring system provides notification of any potential threats. ARM has a cybersecurity policy that covers data collection sites and end-to-end encrypted data transfer to the ADC located at ORNL. The security policy also incorporates disaster recovery plans. Purity detection checkpoints provide real-time notification of potential breaches of any of our systems.

As explained in R9, all data are stored in three different archival systems (ADC storage, ORNL HPSS, and Argonne HPSS), an active process that is already implemented in operations and can be used for swift recovery. In addition, all the systems are backed up daily to ensure protection in the event of an outage.

ARM data services has implemented a detailed cyber security and incident response plan. Because of the security requirements, this plan cannot be shared externally, but the following paragraphs are some excerpts from this document.

The ARM cyber infrastructure resides in three different areas: the ADC at ORNL; the measurement facilities, which include resources at Argonne; and the ORNL Leadership Computing Facility (OLCF). The ORNL Cyber Security Program is responsible for investigating and reporting incidents affecting the infrastructure at the ADC and in the OLCF. The Argonne Cyber Security Program Office is responsible for investigating and reporting incidents affecting the infrastructure at the measurement facilities, as well as all traffic to and from the measurement facilities routed through Argonne.

Triage detection: Odd behavior is to be reported to the proper system administrator immediately. The system administrator will determine if the behavior is the result of a compromise

Impact assessment: The level of impact is determined by the extent of the confidentiality, integrity, and availability of the

systems/services in question. It is also determined by the level of access of a compromised system/service to ARM data, as well as by whether the incident is public facing (e.g., a web page defacement).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: