ARM Data and Computing Capabilities

GIRI PRAKASH

ARM Data Center, Oak Ridge National Laboratory
palanisamyg@ornl.gov

BERAC, April 21, 2023

Comprehensive Sets of Measurements Deployed in Diverse Climate Regimes

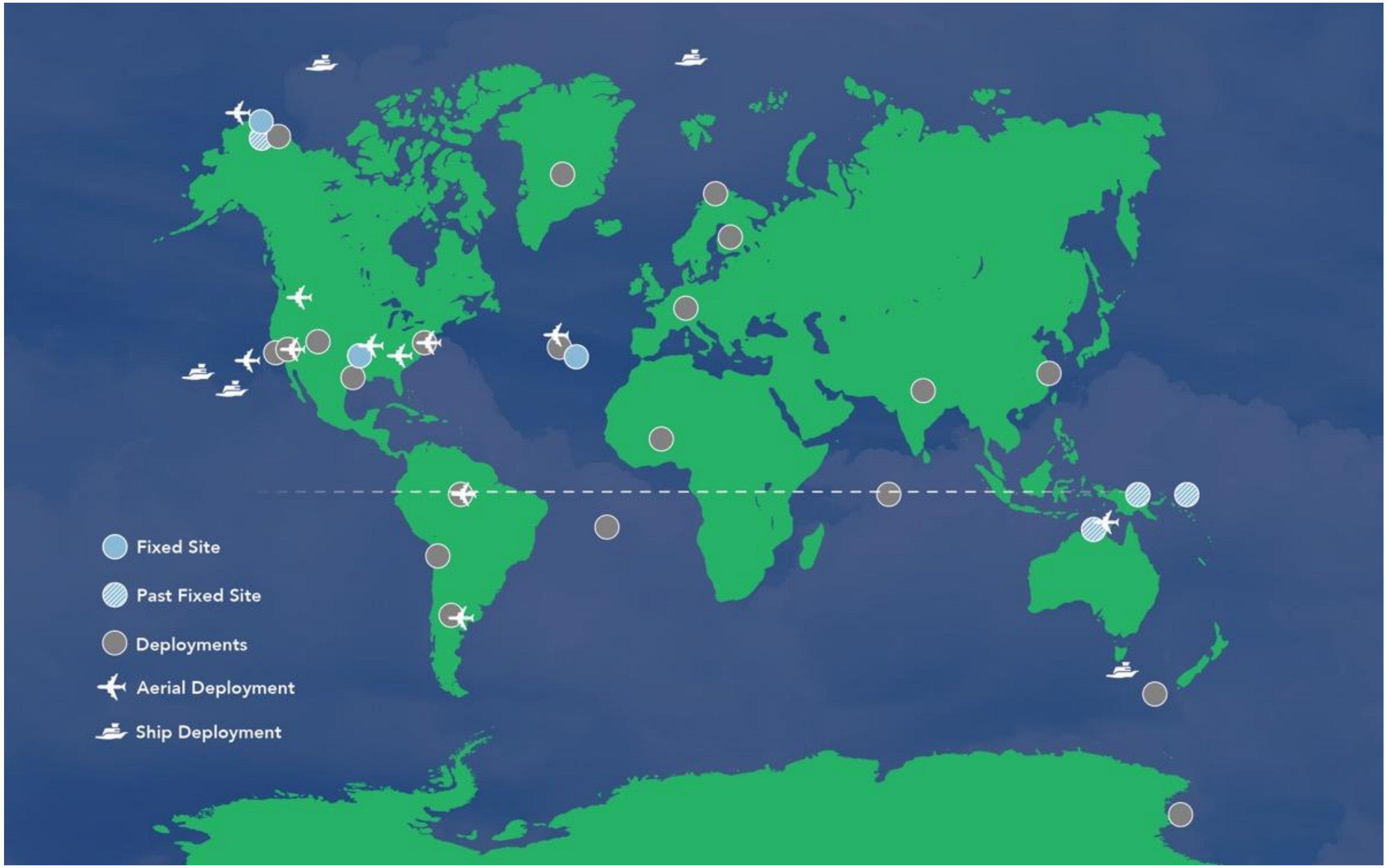
Background atmospheric state

Surface energy balance

Aerosol and hydrometeor profiles

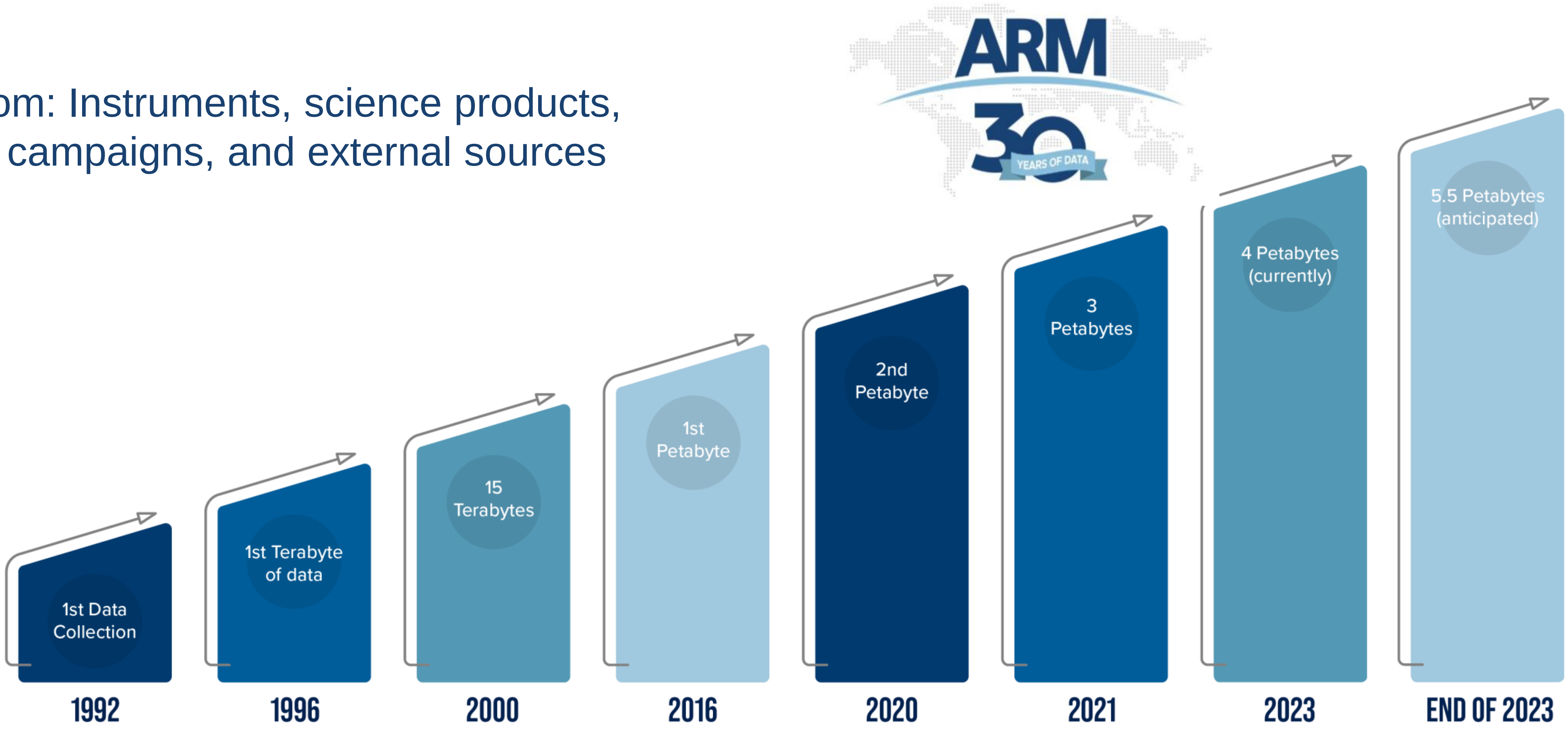
Near-surface aerosol properties

Aerial measurements

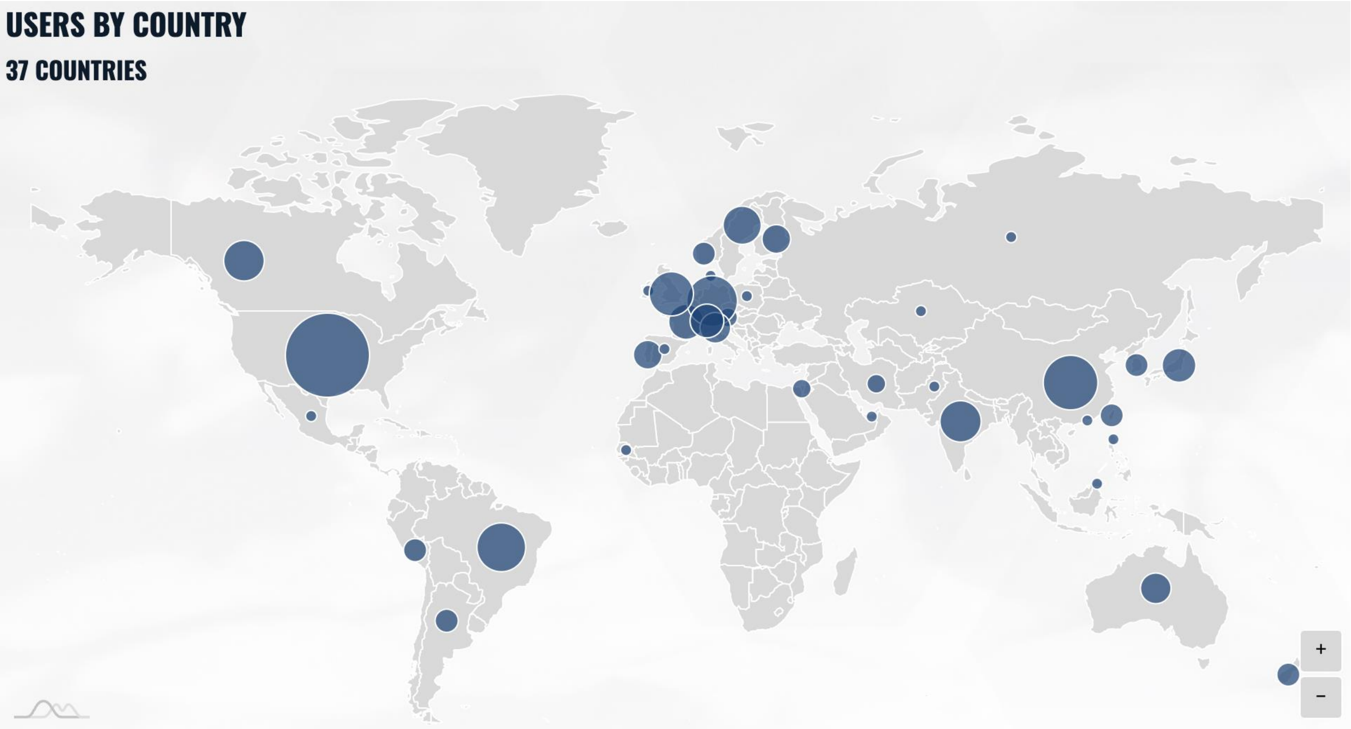


Data and Users At a Glance

- Data From: Instruments, science products, models, campaigns, and external sources



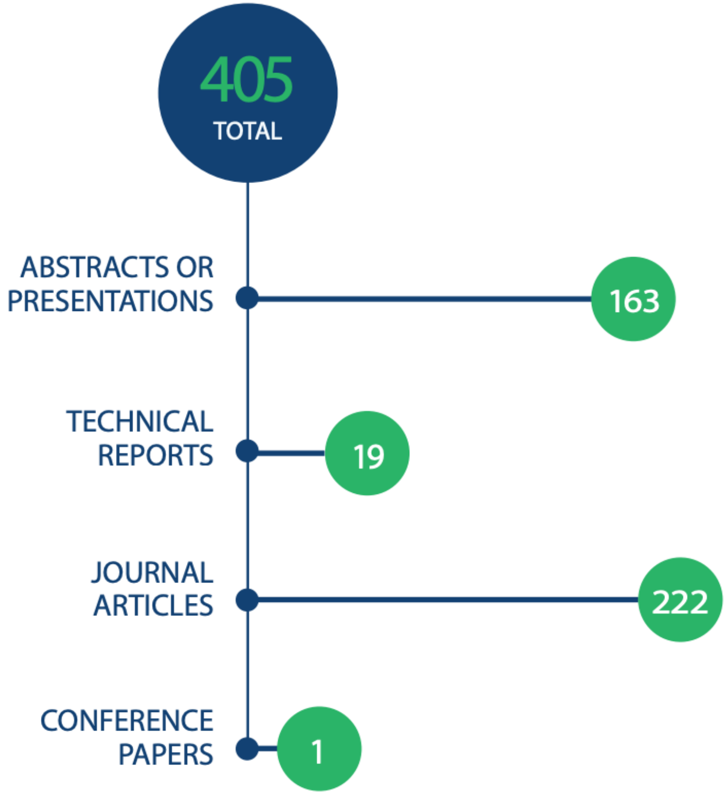
Data and Users At a Glance



USERS BY COUNTRY

37 COUNTRIES

PUBLICATIONS USING ARM*

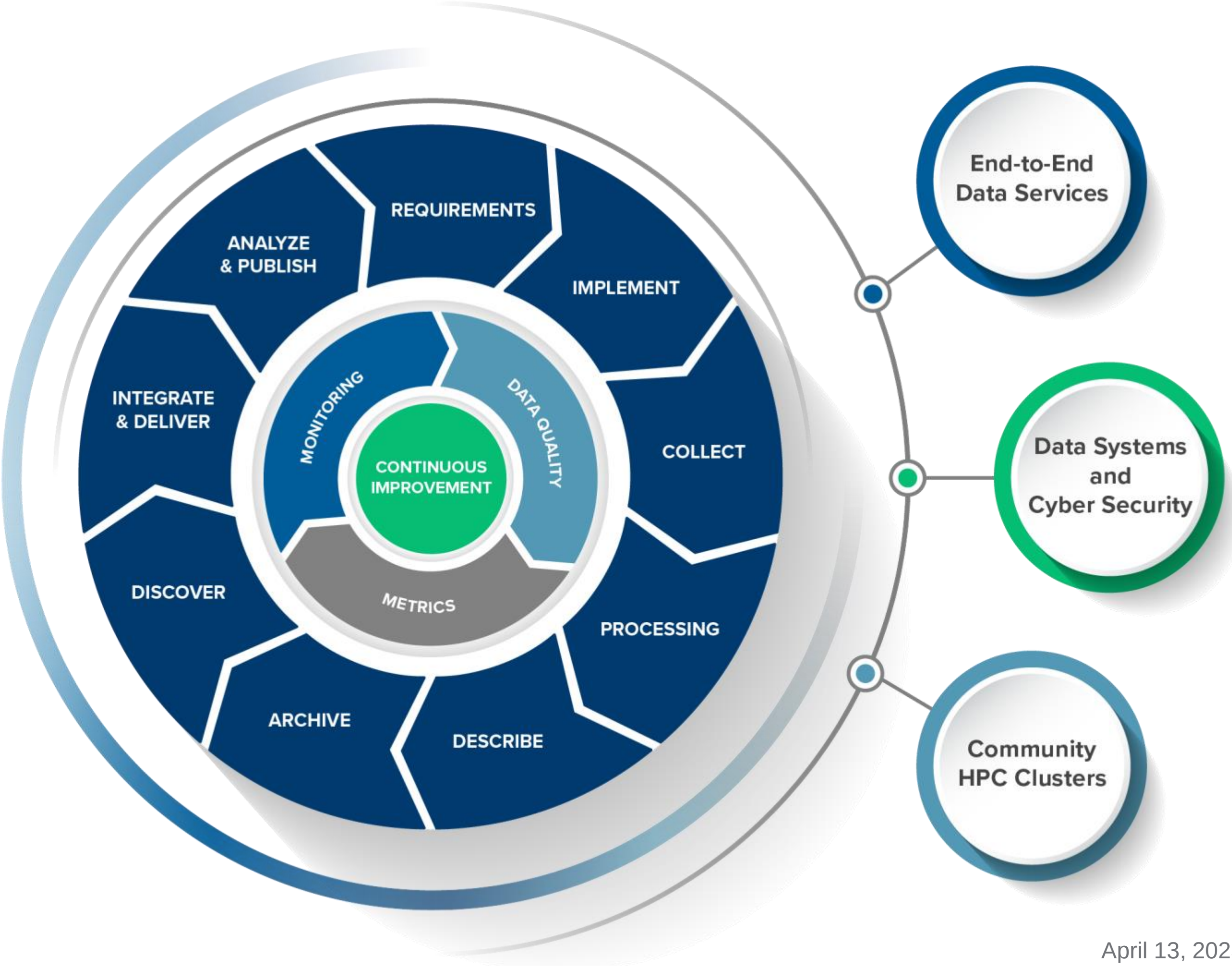


*Publication statistics were collected as of December 2022. Journal article numbers will continue to increase over time.

About The ARM Data Services

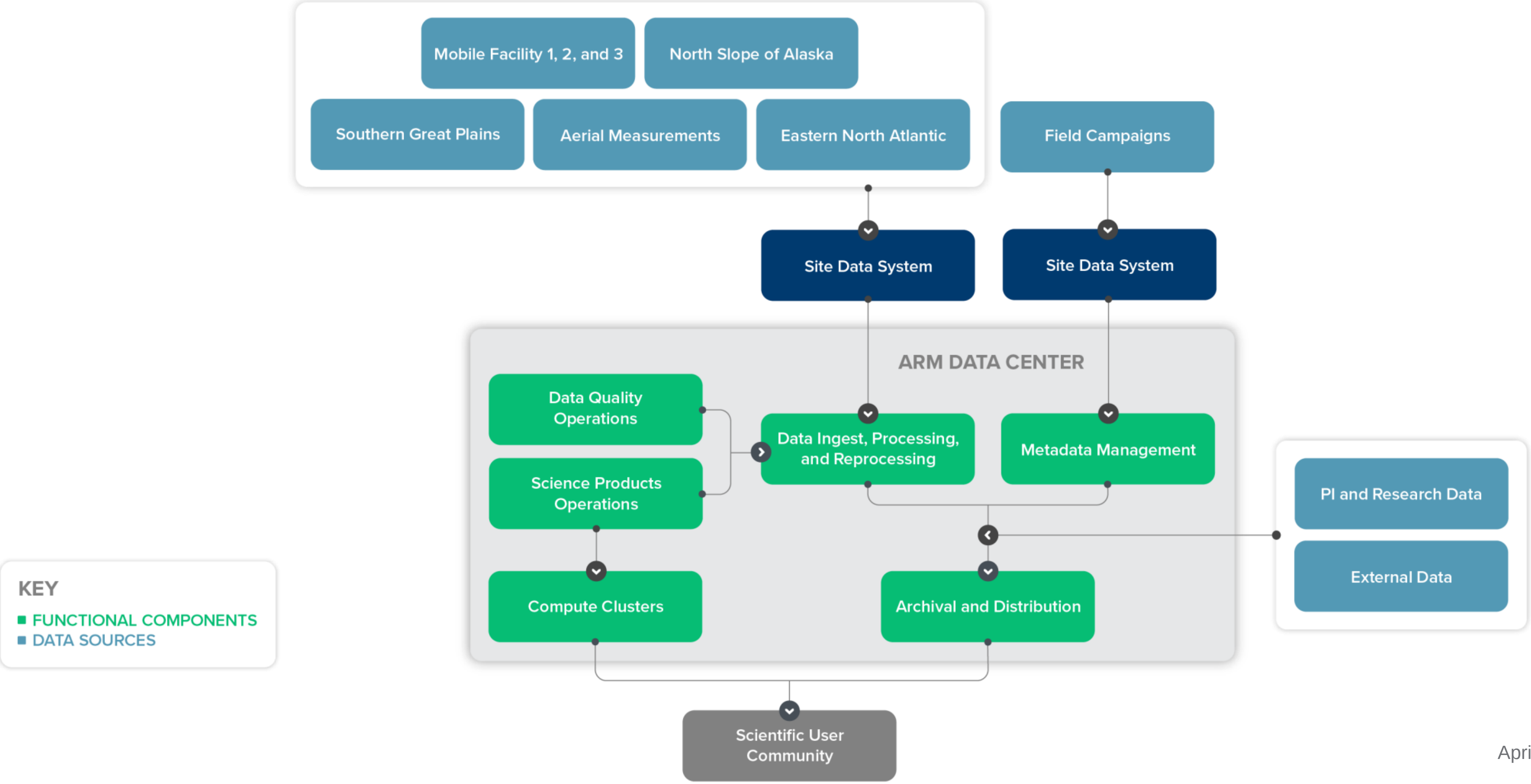
Provides a robust integrated data and computing ecosystem to advance understanding of atmospheric measurements

- Data flow operations and monitoring
- Advanced data collection systems
- High-performance computing (HPC)
- Comprehensive Data Processing
- Data Interoperability:
  - Advanced strategies for utilizing metadata
  - Data Discovery
  - Data workbench
  - FAIR, Standards, and Protocols
- User Management and Citations
- AI-based approach in data management



ARM Data Flow: From Collection to Distribution

- Offers powerful and adaptable infrastructure capabilities to support a wide range of data pipeline requirements, enabling efficient and streamlined processing of data from various sources.



Mobile Facility 1, 2, and 3

North Slope of Alaska

Southern Great Plains

Aerial Measurements

Eastern North Atlantic

Field Campaigns

Site Data System

Site Data System

ARM DATA CENTER

Data Quality Operations

Science Products Operations

Data Ingest, Processing, and Reprocessing

Metadata Management

Compute Clusters

Archival and Distribution

PI and Research Data

External Data

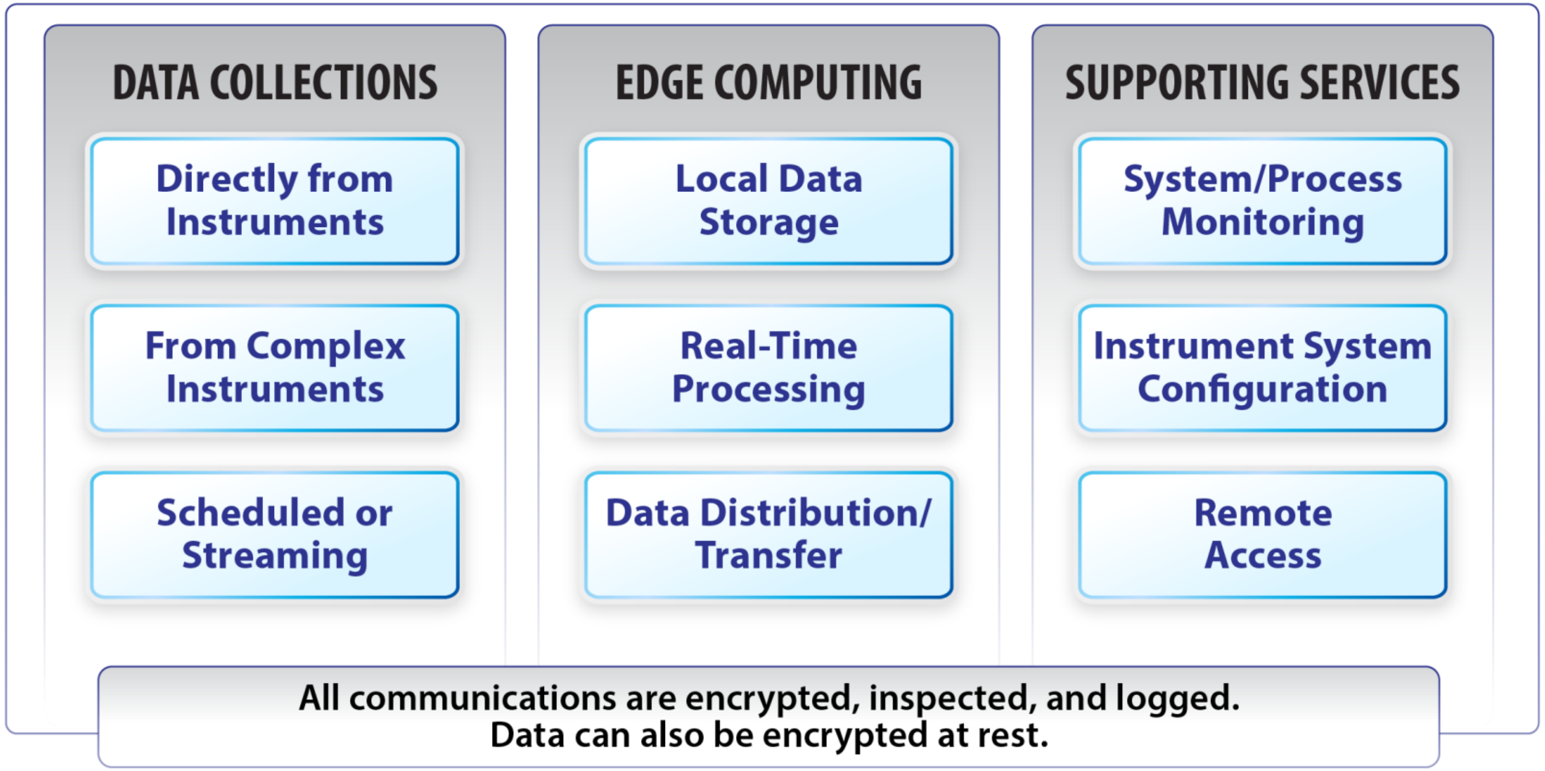
Scientific User Community

KEY

- 🟩 FUNCTIONAL COMPONENTS
- 🟦 DATA SOURCES

Advanced-Data Collection Systems For Next-Gen Sensor Networks

- Scalable data systems with proven hardware and software solutions
- Real-time data access to enable data reduction and edge computing (e.g., Supervised Learning)
- Future development of next-generation instrument computing with Machine Learning



DATA COLLECTIONS

- Directly from Instruments
- From Complex Instruments
- Scheduled or Streaming

EDGE COMPUTING

- Local Data Storage
- Real-Time Processing
- Data Distribution/ Transfer

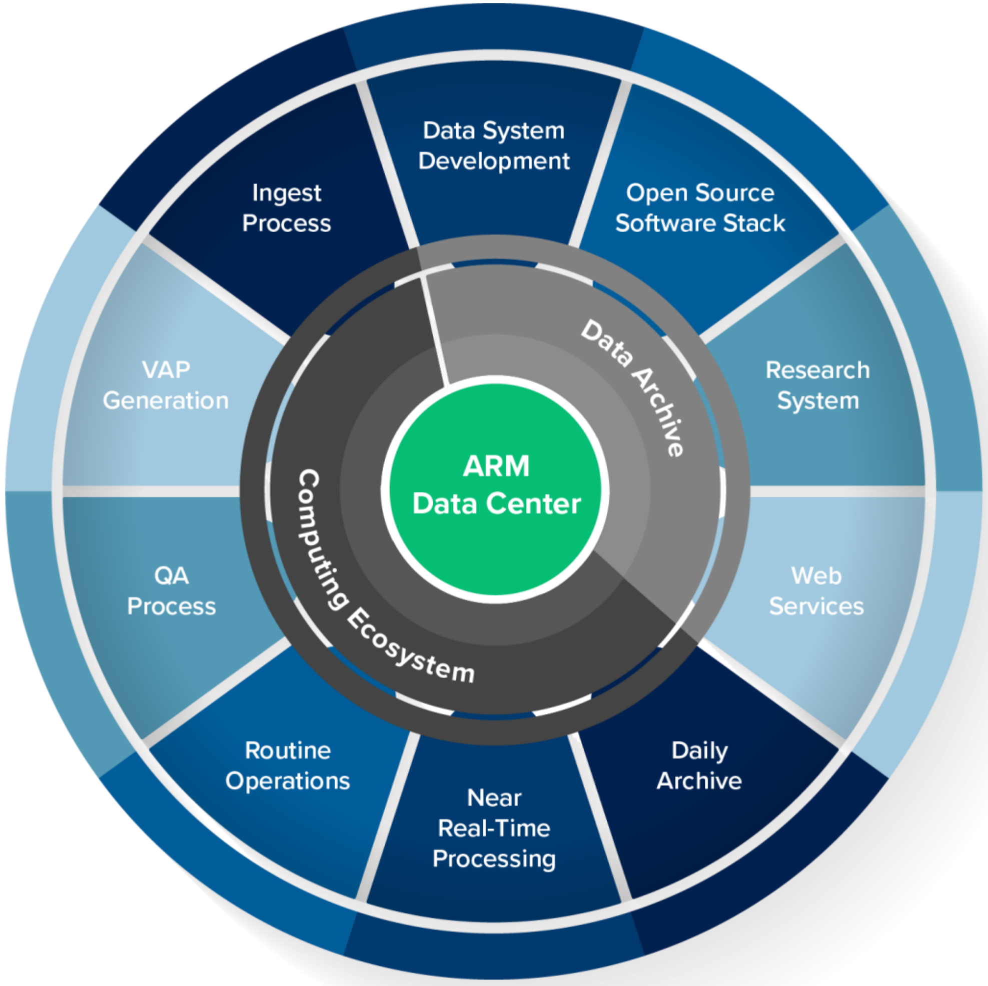
SUPPORTING SERVICES

- System/Process Monitoring
- Instrument System Configuration
- Remote Access

All communications are encrypted, inspected, and logged.
Data can also be encrypted at rest.

Comprehensive Data Processing Capabilities For Effective Data Management

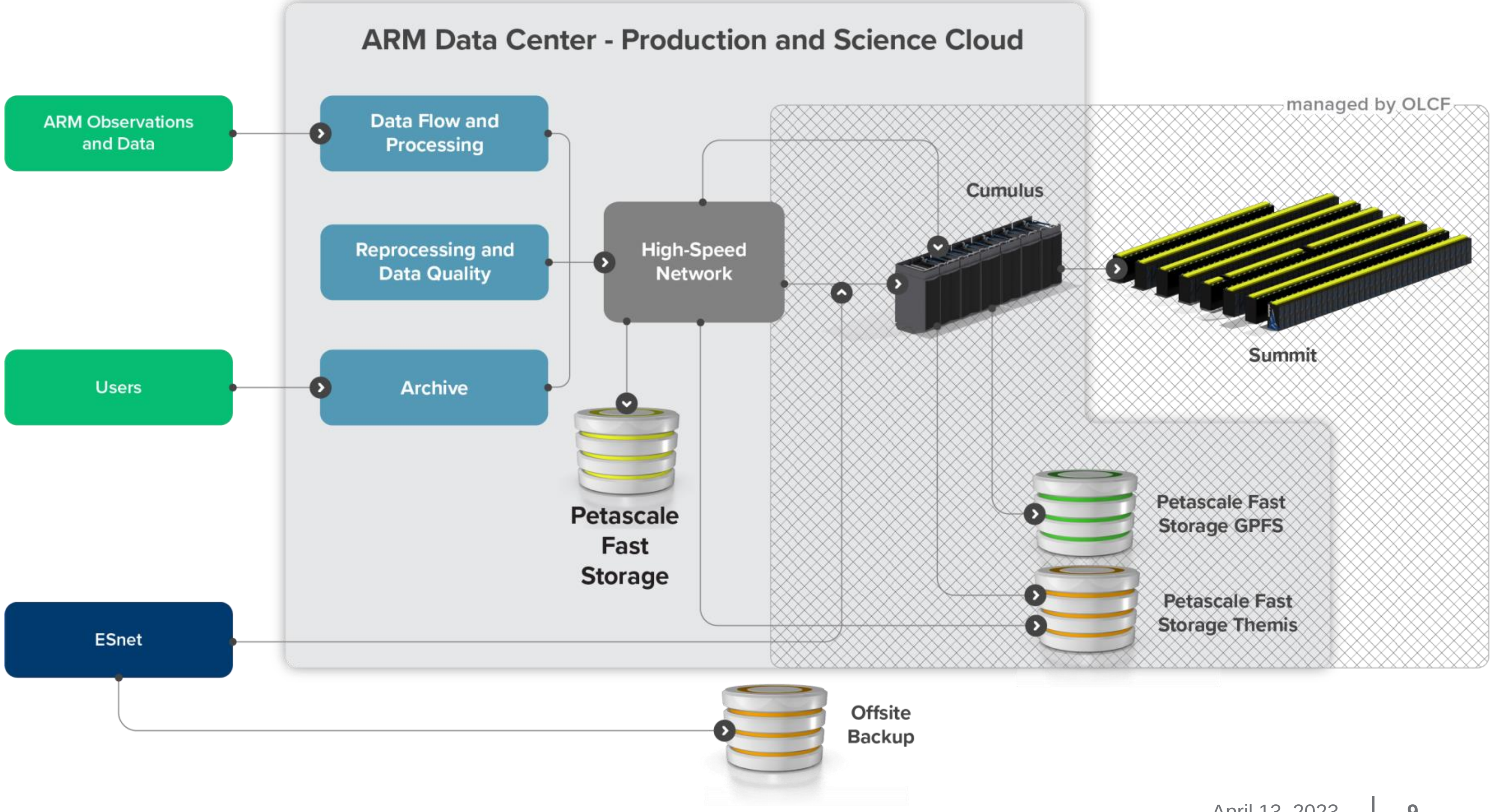
- Efficient flow control for smooth processing
- Thorough data quality review for accuracy and reliability
- Immediate online access to high-demand data streams
- Near-term and long-term reprocessing capabilities
- Comprehensive monitoring of thousands of processing data streams.



Computing Capabilities

ARM Data Center Cyberinfrastructure: Enhancing synergy across DOE computing facilities

- Enables fast parallel processing of:
  - Data ingest operations
  - Complex ARM datastreams (e.g., Radars, Value Added Products)
  - Large Eddy Model Simulations of ARM cases (LASSO)
  - Data analysis using large volumes of ARM data

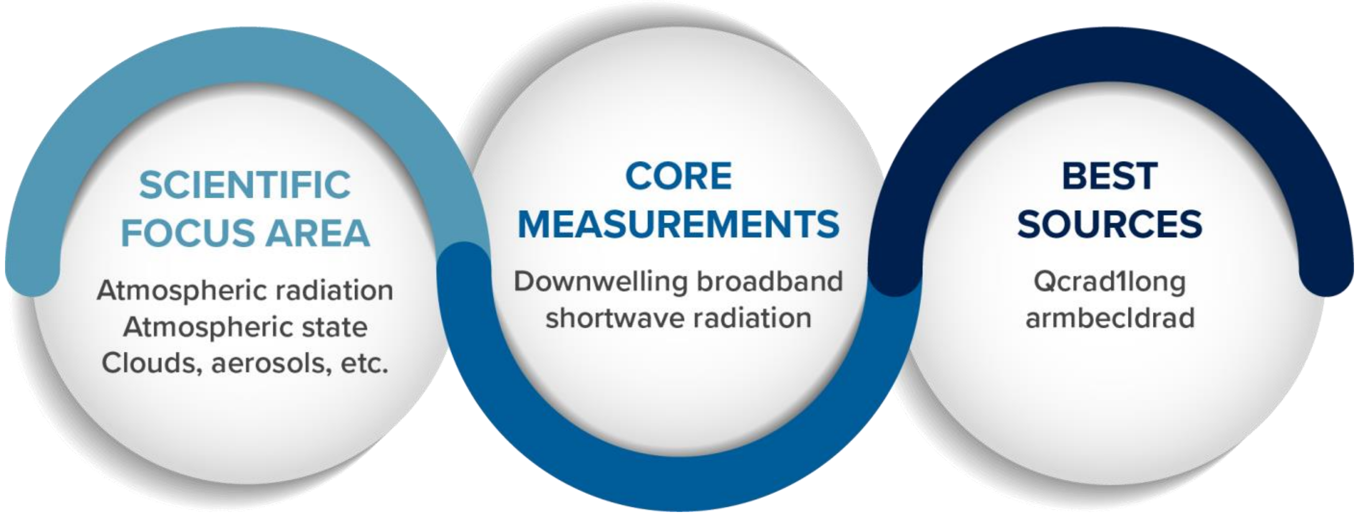


ARM Data Center - Production and Science Cloud
ARM Observations and Data
Users
ESnet
Data Flow and Processing
Reprocessing and Data Quality
Archive
High-Speed Network
Petascale Fast Storage
managed by OLCF
Cumulus
Summit
Petascale Fast Storage GPFS
Petascale Fast Storage Themis
Offsite Backup

Selecting Quality Data Sources: Harnessing the Power of Rich Metadata

Over 11,000 Data products from 450+ instruments, science products, and model simulations

- Robust metadata workflow system effectively used for operations, discovery, and data interoperability
- Recommends best data sources for the core measurements (i.e., Data Epoch)
- Semi-automated process includes input from subject matter experts



SCIENTIFIC FOCUS AREA

Atmospheric radiation
Atmospheric state
Clouds, aerosols, etc.

CORE MEASUREMENTS

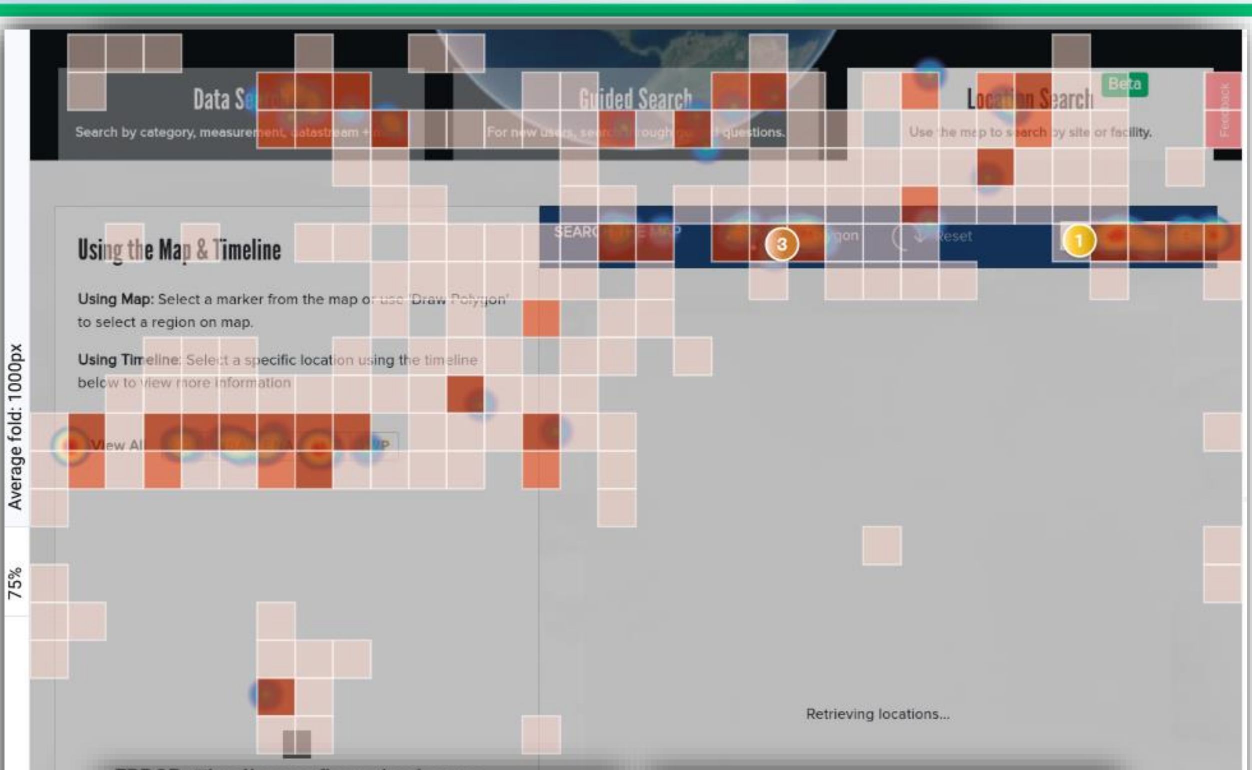
Downwelling broadband
shortwave radiation

BEST SOURCES

Qcrad1long
armbecldrad

Advanced Data Discovery: Leveraging Modern Architecture and Search Capabilities

- User-centric design and improvements using modern software architecture with Continuous Integration and Deployment (CI/CD)
- Intelligent search capabilities based on the actual data, guided search based on user experience
- Recommendations, data tagging based on epochs or golden periods
- Near real-time access via secured webservices (API access)
- Customized interface for ARM high-resolution model simulations



Search by category, measurement, datastream + ...

For new users, search through guided questions.

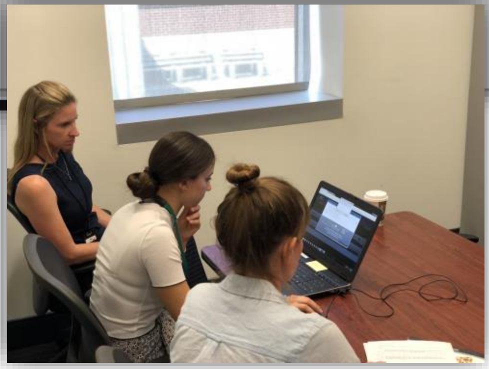
Use the map to search by site or facility.

Using the Map & Timeline

Using Map: Select a marker from the map or use 'Draw Polygon' to select a region on map.

Using Timeline: Select a specific location using the timeline below to view more information

Retrieving locations...



Data Workbench: Enabling Data Interoperability

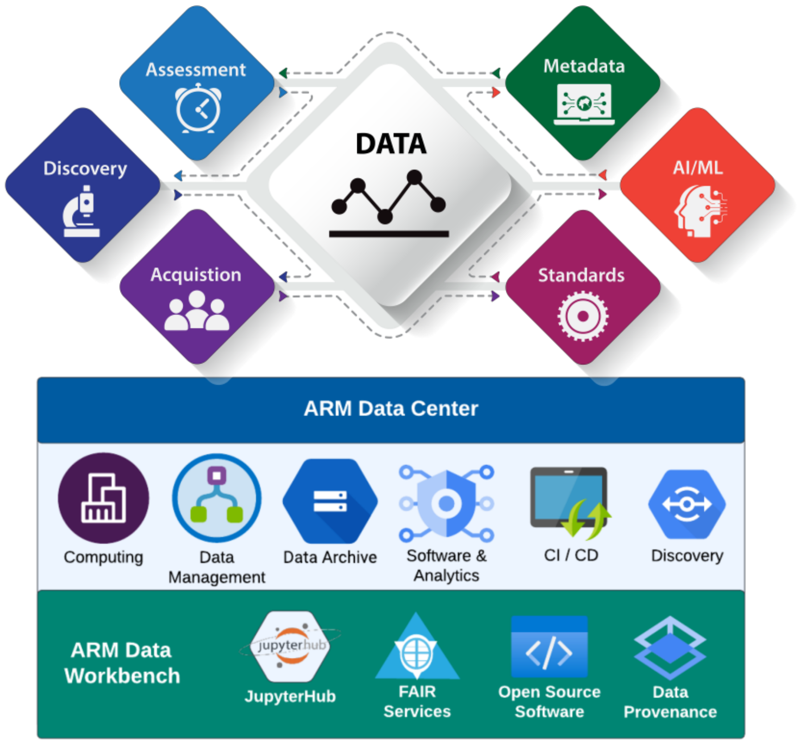
- Aims to achieve transformative knowledge discovery by providing modular computing, data, and software capabilities

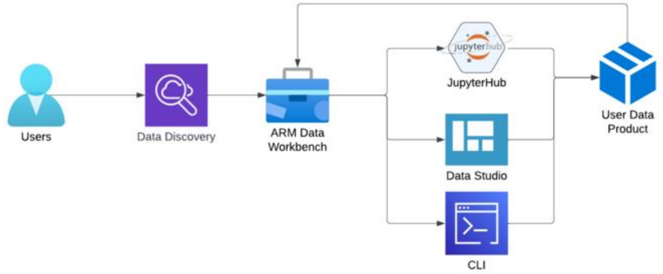
- Facilitate easier interaction with ARM data and enable interoperability with other data sources

  - Provide a collaborative and dynamic computation environment for data analysis, scientific computing, and machine learning (e.g., JupyterHub)

  - Facilitate data access to external datasets (e.g., weather radar, satellite, model data, etc.)

- Enable FAIR-based access to ARM data and computing for initiatives such as AI4ESP





FAIRness Assessment and Community Engagement

PuRe Data Resources

CORE TRUST SEAL

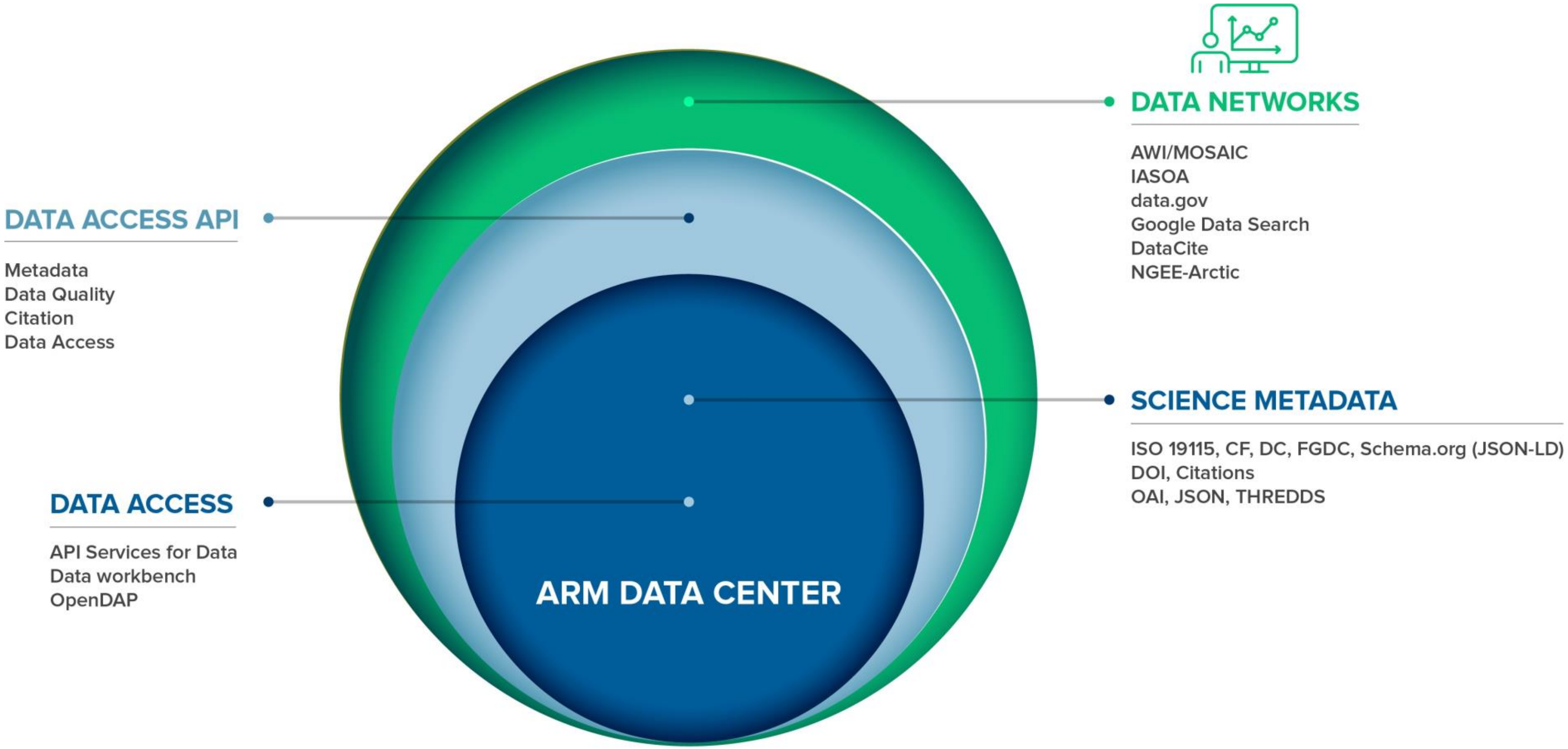
WORLD DATA SYSTEM

Findable Accessible Interoperable Reusable

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

- Review of data management capabilities and obtaining certifications
- Continuous collaboration with broader data networks
- Active contribution to national and international working groups

Putting FAIR Principles into Practice: Standards and Protocols in Data Interoperability



DATA NETWORKS

AWI/MOSAIC
IASOA
data.gov
Google Data Search
DataCite
NGEE-Arctic

DATA ACCESS API

Metadata
Data Quality
Citation
Data Access

SCIENCE METADATA

ISO 19115, CF, DC, FGDC, Schema.org (JSON-LD)
DOI, Citations
OAI, JSON, THREDDS

DATA ACCESS

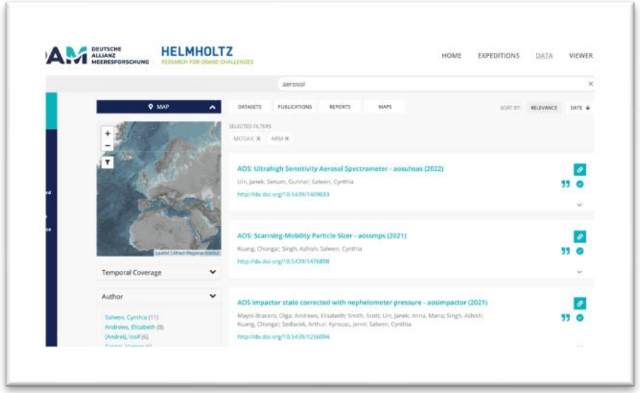
API Services for Data
Data workbench
OpenDAP

ARM DATA CENTER

Expanding the Reach of ARM Data: Data sharing Examples

- Data Access:
  - Ensure the latest version of data are available for users
  - Data endpoints are provided in the metadata
  - Direct access via API-based services (live data service, Globus, OpenDAP, and JupyterHub)
- Provide access to data quality, plots, and other ancillary details
- Options for users to get notified of any data quality changes or new data versions
- Interoperability:
  - ARM Data is currently discoverable in partner portals
  - External data are shared through the ARM discovery interface
  - Currently in discussion with BER data centers such as EMSL, ESGF, and Ameriflux

MOSAIC Data Portal



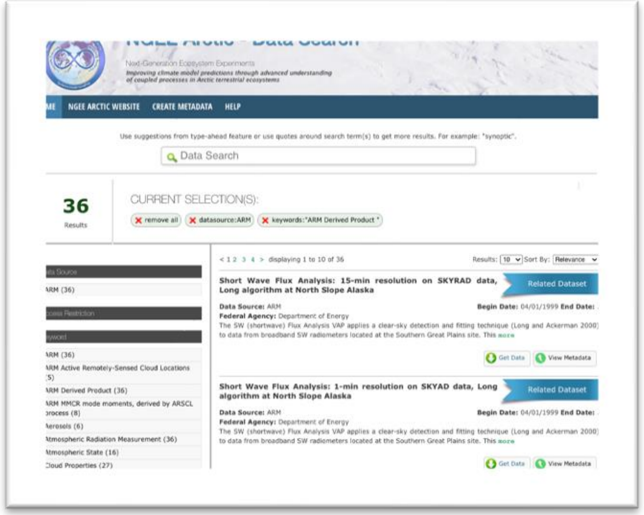
AOS: Ultrahigh Sensitivity Aerosol Spectrometer - aosuhsas (2022)
Uin, Janek; Senum, Gunnar; Salwen, Cynthia
http://dx.doi.org/10.5439/1409033

AOS: Scanning-Mobility Particle Sizer - aossmps (2021)
Kuang, Chongai; Singh, Ashish; Salwen, Cynthia
http://dx.doi.org/10.5439/1476898

AOS impactor state corrected with nephelometer pressure - aosimpactor (2021)
Mayol-Bracero, Olga; Andrews, Elisabeth; Smith, Scott; Uin, Janek; Anna, Maria; Singh, Ashish; Kuang, Chongai; Sedlacek, Arthur; Kyrouac, Jenni; Salwen, Cynthia
http://dx.doi.org/10.5439/1256094

Salwen, Cynthia (11)
Andrews, Elisabeth (8)
(Andrei), Iosif (6)

NGEE-Arctic Data Portal


Next-Generation Ecosystem Experiments
Improving climate model predictions through advanced understanding of coupled processes in Arctic terrestrial ecosystems
NGEE ARCTIC WEBSITE
CREATE METADATA
HELP
Data Search
Related Dataset
Get Data
View Metadata

Use suggestions from type-ahead feature or use quotes around search term(s) to get more results. For example: "synoptic".

36 Results

CURRENT SELECTION(S): remove all | datasource:ARM | keywords:"ARM Derived Product"

< 1 2 3 4 > displaying 1 to 10 of 36 Results: 10 Sort By: Relevance

ARM (36)
ARM Active Remotely-Sensed Cloud Locations (5)
ARM Derived Product (36)
ARM MMCR mode moments, derived by ARSCL process (8)
Aerosols (6)
Atmospheric Radiation Measurement (36)
Atmospheric State (16)
Cloud Properties (27)

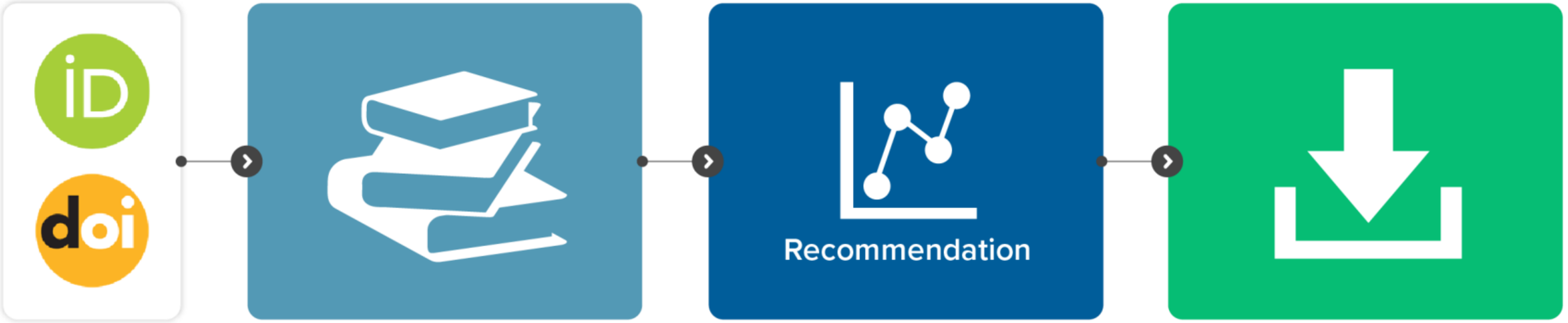
Short Wave Flux Analysis: 15-min resolution on SKYRAD data, Long algorithm at North Slope Alaska
Data Source: ARM **Begin Date:** 04/01/1999 **End Date:**
Federal Agency: Department of Energy
The SW (shortwave) Flux Analysis VAP applies a clear-sky detection and fitting technique (Long and Ackerman 2000) to data from broadband SW radiometers located at the Southern Great Plains site. This more

Short Wave Flux Analysis: 1-min resolution on SKYAD data, Long algorithm at North Slope Alaska
Data Source: ARM **Begin Date:** 04/01/1999 **End Date:**
Federal Agency: Department of Energy
The SW (shortwave) Flux Analysis VAP applies a clear-sky detection and fitting technique (Long and Ackerman 2000) to data from broadband SW radiometers located at the Southern Great Plains site. This more

Maximizing User Management and Data Citations

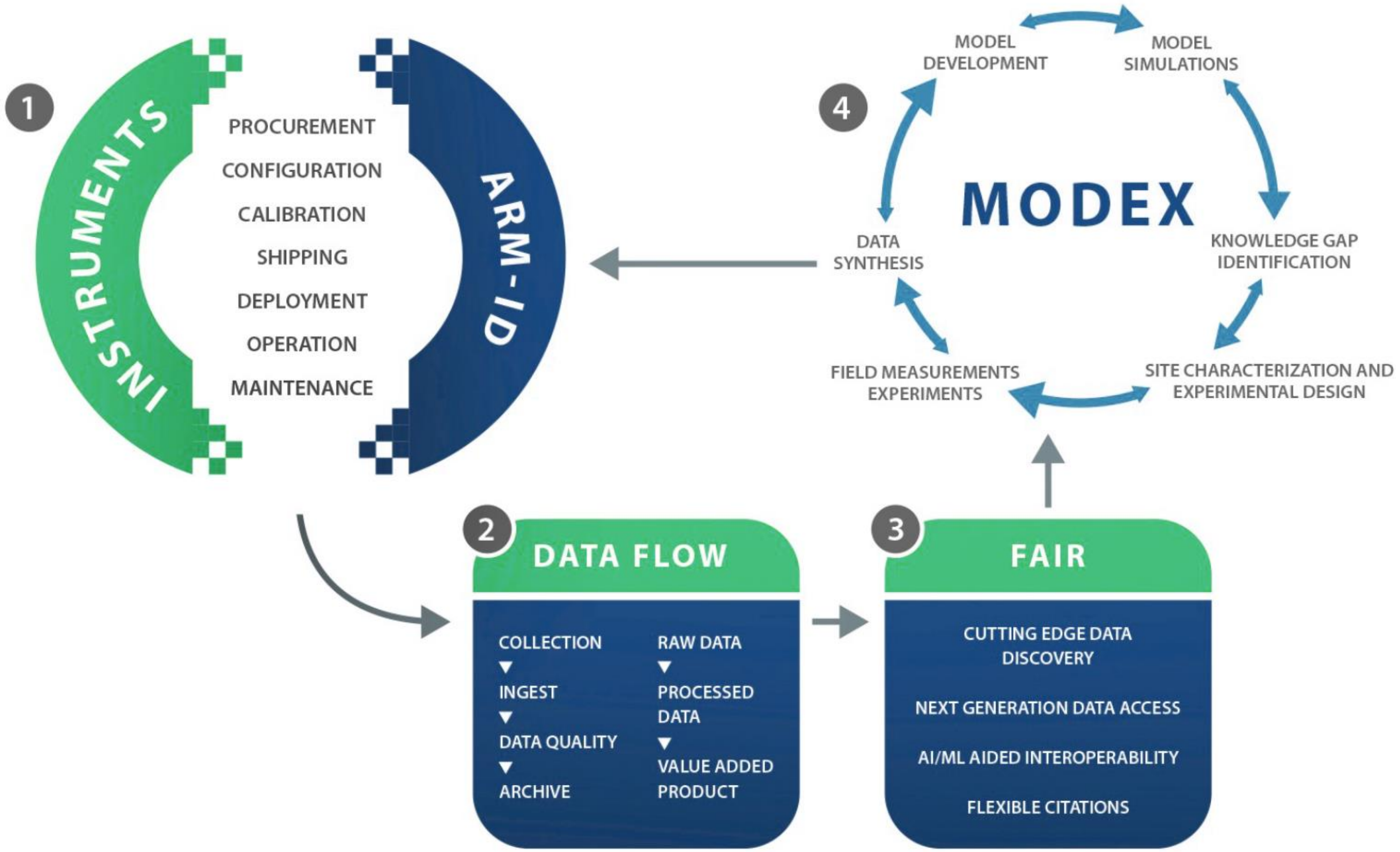
- Integrating ORCID with other user metrics improves the program's ability to manage the quality of user details and metrics preparation

- Opportunities exist to improve user experience using AI/ML techniques
  - Discover relationships between ORCID identifiers, users, publications, data, metadata etc. Then use these relationships to improve the user experience with finding and using ARM Data



Looking Ahead: Unlocking the Power of Data. The Role of AI in Enhancing Observational Data Centers

- Enabling interdisciplinary research through modernization of data pipelines from collection to distribution using AI-based approaches
- Near real-time data analysis and data collection configurations using edge computing
- Developing and extending community-based standards between data repositories and AI models
- Data tagging to identify benchmarking/training datasets

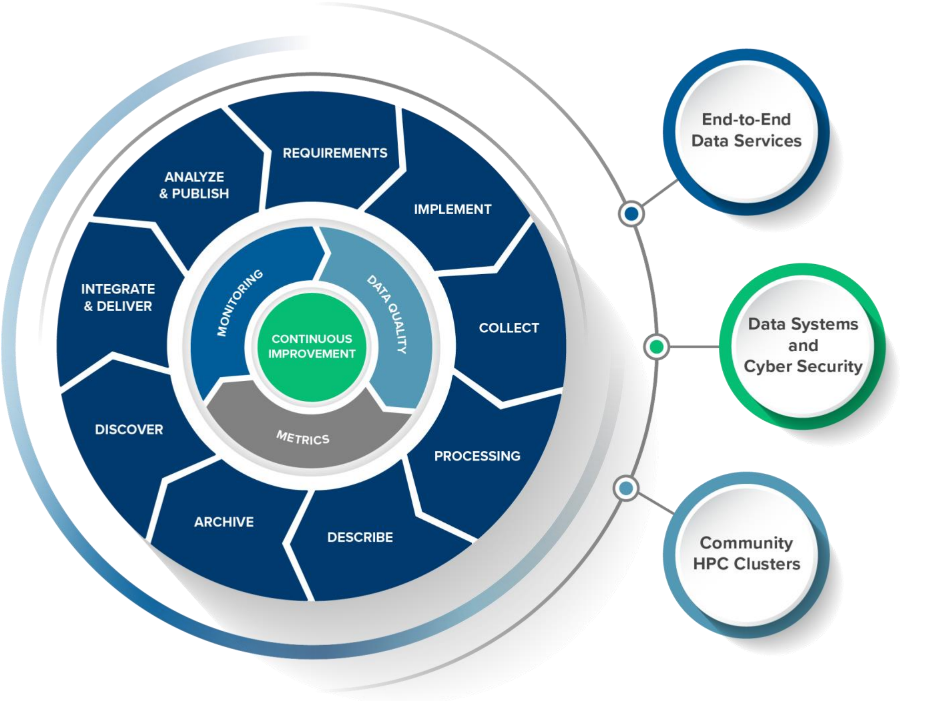


1
INSTRUMENTS
PROCUREMENT
CONFIGURATION
CALIBRATION
SHIPPING
DEPLOYMENT
OPERATION
MAINTENANCE
ARM-ID
4
MODEX
MODEL DEVELOPMENT
MODEL SIMULATIONS
KNOWLEDGE GAP IDENTIFICATION
SITE CHARACTERIZATION AND EXPERIMENTAL DESIGN
FIELD MEASUREMENTS EXPERIMENTS
DATA SYNTHESIS
2
DATA FLOW
COLLECTION ▼ INGEST ▼ DATA QUALITY ▼ ARCHIVE
RAW DATA ▼ PROCESSED DATA ▼ VALUE ADDED PRODUCT
3
FAIR
CUTTING EDGE DATA DISCOVERY
NEXT GENERATION DATA ACCESS
AI/ML AIDED INTEROPERABILITY
FLEXIBLE CITATIONS

Summary

ARM Data Services

- Provides robust data collection, processing, archival, and distribution capabilities
- Enables unified data, computing, and software ecosystem for scientists and facility operations
- Empowers data interoperability with broader scientific data networks by putting FAIR principles into practice.


REQUIREMENTS
IMPLEMENT
COLLECT
PROCESSING
DESCRIBE
ARCHIVE
DISCOVER
INTEGRATE & DELIVER
ANALYZE & PUBLISH
MONITORING
DATA QUALITY
METRICS
CONTINUOUS IMPROVEMENT
End-to-End Data Services
Data Systems and Cyber Security
Community HPC Clusters