Facilities/Instruments Data Case Studies Template

Galen Shipman (ORNL), Kerstin Kleese-Van Dam (PNNL), Barbara Jennings (SNL) and Eric Ackerman (SNL)

This case study was derived from previous requirement gathering activities and public reports including the ARM ESnet requirements document 2012 and the report from the BERAC advisory committee "BER Virtual Laboratory – Innovative Framework for Biological and Environmental Grand Challenges". Furthermore specific requirements from the perspective of different ARM user groups were added to the document throughout. More detailing requirements analysis with BER facilities and end-users would be beneficial.

1.1 Background

Please provide a brief description of the science, facility/instrument and user communities. For what purpose are you generating and collecting the data?

The Atmospheric Radiation Measurement (ARM) Climate Research Facility is a long-term measurement facility funded by CESD of BER in DOE that focuses on measuring:

- Cloud properties: microphysics (phases of water), optical properties, and patterns of occurrence
- Aerosol properties: size, chemistry, optical, and generation and decay pathways
- Cloud and aerosol interactions: absorption of aerosols by clouds and cloud formation triggered by aerosols
- Sunlight energy "fate": radiative flux transfer, heating rate profiles, components of reflected and absorbed radiant energy, direct and diffuse light
- Atmospheric state: profiles of temperature, water vapor, wind, and aerosols
- Surface properties: Surface fluxes, soil conditions

The ARM climate research facility is building a "climatology" (multi-year record) of these measurements that are related to cloud formation, sunlight energy fate, and aerosol formation / decay, and aerosol interactions with clouds. The measurements are used to improve parameters that represent these processes in global circulation models (GCMs). The GCM models are used for the prediction of future climate patterns. Parameters for cloud formation, sunlight energy fate, and aerosol interactions are thought to be the source of the largest uncertainties in these models and long-term climate forecasts.

The need for data fusion is one of the challenges that face researchers in the ARM program. Monitoring data is collected by a number of various entities each with their unique data resource. The ability to easily access such disparate resources could greatly enhance the relevance that the individual programs have on each other. Moreover, the data-fusion challenge must extend far beyond the ARM program to produce reliable climate models. For example, additional agencies collect similar complementary data,

and integration with other sciences, such as biological and geological, are also essential for worthwhile models.

1.1.1 Instruments and Facilities

Please provide an overview description of your facility and instruments incl. compute, storage, and network capabilities, describe your service provision in terms of hardware, data acquisition, data management, software infrastructure and applications.

ARM field sites are located in Oklahoma / Kansas, North Slope of Alaska, Tropical Western Pacific (Manus and Nauru Islands and Darwin Australia). The ARM mobile facility is currently located in Cape Cod, and soon to be in Brazil. Formerly it was in India, Azores, China, Germany, Niger, Africa, and Coastal California. A second, more modular, Mobile facility is operating on a cargo ship in the Pacific and was previously in Maldives and Colorado. This second Mobile facility is designed to be ship-based, as needed. Two additional sites are being developed to bring online in 2014 in the Azores and North Slope of Alaska. As a user facility, ARM regularly has field campaigns collocated with existing sites that involve collaborations with the entire atmospheric community and their instruments. The program also has ties with NOAA, NASA, and ECMWF (European Centre for Medium-Range Weather Forecasts), but few interactions with DoD.

Field data systems are located at each of the field and mobile sites. Data systems with facility-wide functions are located at BNL and ORNL. A distinct aspect of the ARM data collection is that it is continuous and has essentially the same parameters for its entire history. Most other studies in these aspects of atmospheric science include only short-term case studies of only a few weeks or months.

Most of the Climate modeling is done using the Community Climate System Model (CESM).

At SNL researchers want to understand the interplay of biosphere and climate. They need to predict the effects on arid land ecosystems. To do this they need to integrate, manage and utilize biological, geological and atmospheric data. They need to move large scale gnome data into climate models. This effort requires the collaboration of BER, LBL, JGI UNM and SNL.

The users of the ARM data and network resources include: ARM facility personnel (for initial data collection, internal transfer, processing, and storage) and the research community (for access / download / use of documentation and data). The user community is mostly in the US, but is also globally distributed. Users can be divided into the following categories: working at DOE facilities; not working at DOE but within the US; located at universities; and persons from foreign countries.

1.1.2 Data-Driven Process of Science

Please provide a description of the process by which data is generated, collected, and processed; are the raw data transformed, restructured, or reduced in volume at the your facility/instrument site? How do the scientists use the instruments and facilities for knowledge discovery? Please outline which processes are carried out at the facility/instrument site and what your users are doing after they used your

facility/instrument site. Do you explicitly design your experiments, either through simulations, or based on previously-run experiments? Do you use the results of analysis for making decisions? If so, provide details.

The process of science can be split into two distinct steps, scientific work within the facility and further scientific exploration of the results of the facilities work by its wider user community.

The ARM data management facility (DMF) at PNNL collects, calibrates, cleans, quality checks, annotates and transforms (into ARM NetCDF) all observational results from over 2000 data streams before storing them in the ARM archive at ORNL for user access. ARM has been actively curating its data since its inception over 30 years ago, frequently re-processing and adapting data based on new insights and standards (data format/metadata/QA). In collaboration with the domain scientists and instrument mentors the data management facilities team develop new higher level data products based on the ARM collected data, and data products from other NASA, NCAR, ECMWF etc. The data management facilities team develops data processing pipelines for each new product, which are then operated either for the duration of the experiment or in perpetuity for long range observations.

ARM users have typically downloaded data sets relevant to their research from the Archive and performed further analysis, integration and refinement at their local institution. Growing data volumes, such as those expected from the new scanning Radars, as well as the need to access longer time ranges threaten continued functioning of this traditional approach to access data.

ARM has several developments that may reduce the data volumes to be transferred over the network. Researchers often only need a small portion of a data set for their work, and ARM is developing tools to extract and subset data in ways that make small portions of the large sets of ARM data more easily accessible to scientists. A recent survey of ARM scientists showed that a majority would have reservations about downloading more than 100GB of data per task. Nonetheless, some analyses are likely to require far higher levels of data.

For integrated ecosystem experiments such as the Next-Generation Ecosystem Experiment (NGEE-Arctic), the workflow can be quite complex requiring integration of data from observation, laboratory experiments, and simulation studies. Projects such as NGEE-Arctic aim to understand ecosystem processes that are currently poorly understood, such as permafrost in the Arctic. The workflows within this project vary and may begin with using regional models coupled with observations in order to identify areas in the Arctic where ecosystem processes are not well understood. Sites within the Arctic are then selected for a field campaign in which physical experiments are conducted to understand processes that occur at varying spatial and temporal resolutions and among different system organizations. Experiments range from microbial studies within a laboratory environment to field studies in which the permafrost is forcibly heated while embedded sensors provide critical data on soil composition, methane release, etc. Scientists conducting these studies use a variety of tools for analysis from custom scripts, to community-developed software (such as UV-CDAT, NCL, Ferret), to commercial packages (such as IDL and Matlab). It is important to note what distinguishes these activities from those of the traditional climate modeler discussed above: the diversity of the data collected (from text files, to

simple spreadsheet data, to databases, to climate model output, to biological data on carbon cycling, etc.) and the consequent analysis and integration of data at multiple spatial and temporal scales (from meso-scale observations to regional and even global models).

In order to accelerate the understating in the biological contribution to climate-model development there is a need for combined data from high-throughput omics methods; e.g., genomic, transcriptomic, metabolomics, and proteomic. Generation and analyses of these prodigious quantities of omics data is a tremendous computational challenge that is still in its infancy. This data fusion is a need expressed by the climatologists at Sandia. There is a need to integrate, manage and utilize biological, geological and atmospheric data. Integration of joint efforts will enhance collaboration among partners and dramatically increase the quantities of useful data accessible while greatly improving resulting models.

1.1.3 Tools and Resources Used

Is simulation data used together with the measured data for verification and/or validation? Which mathematical, data management and visualization tools are used for analysis, data understanding, and quality assurance? Do you use any workflow tools? What are the data or file formats currently used? Are there issues related to uncertainty in extracting the science from the data? Do you use tools for uncertainty quantification?

Many of the tools used for data stewardship are based on common technologies such as FTP for data collection, relational databases for metadata storage, file systems and tape archives (HPSS) for data storage. Other tools, in particular, QA/QC workflows are based on custom software solutions but are well established and leverage community standards for data models and formats. Data portal technologies for sharing and "summary analysis" i.e. simple 2-d plots are based on well-established technologies such as Mercury for data discovery and other open-source web technologies. Some of the data processing work leverages cloud computing technologies but the overall analysis workflows for the various data products developed via the ARM program are based on processes that are distinct for specific data products. As the data sizes in the ARM program continue to grow with the deployment of 3D scanning radars, producing about 15-80GB/day per unit, users of the ARM facility require new technologies for data analysis. In response to this increase in data rate and data sizes the ARM facility at ORNL has deployed an ARM data cluster allowing users to access tools such as Visit, ParaView, R, IDL, etc. and the requisite hardware to run these tools.

In general, many climate projects (such as ARM, CDIAC, and others) rely upon diverse toolsets and technologies for data management—including Drupal for online content and ORNL's Metadata Clearinghouse for metadata harvesting and search and discovery. Major data centers (such as ARM) have developed robust automated workflow composition and execution for QA/QC of the data that they collect (such as ADI for ARM). These data centers provide reports along with the data they collect describing any noted QA/QC concern with the data.

There is a need for BER's Environmental Division to create its version of the Biological Division's KBase and then integrate both of them to attack climate modeling problems; *i.e.,* a Biological and Environmental KBase. Lessons already learned establishing the biological KBase could shorten

implementation and integration with the environmental KBase. Access to this integrated data would greatly enhance the ability of Climate researches at Sandia to do make reliable long-term predictions. Issues such as environmental tipping points and system resiliencies could then be addressed, particularly in terms of perturbations contributed by the microbial and plant worlds.

1.1.4 Analysis Software

What kind of analysis software is used at the facility/instrument site – community-wide effort, commercial, instrument-specific, user-specific? What is the purpose of performing analysis at the facility/instrument site? What are response time requirements - real time, near real time, post experiment? What analysis is done at the user site and what are the requirements for these analyses? What machine does the analysis run on - instrument workstation, small cluster, HPC system, grid computing resources, cloud computing? Are there science questions you would like to address, but cannot due to lack of appropriate software?

The ARM facility uses a variety of tools for data conversion, transformation and integration to develop data products for the ARM user community, hereby the newly developed ARM Data Integration Framework (ADI) helps to reuse common components and to create operational workflows for the data product generation from both common components and specific analytical algorithms. Each of the ~4000 ARM data products has its own distinct data processing pipeline. The ARM facility makes available a variety of tools to the community for subsequent reanalysis, integration, etc. of the data products that they provide. Users of ARM data products may use commercial software such as MS Excel, IDL, Matlab, to community software such as Visit, R, to custom software and scripts for analysis. Many analysis tasks are conducted on users desktops, large-scale data products (radar/lidar) may require mid-scale compute resources, this is new for the ARM community as traditionally the data sets were small enough to manage on a workstation. An expected outcome is that considerable new open source software will be developed and refined by disparate scientists exploiting the available integrated data while rapidly sharing their results.

1.1.5 Collaboration

Do you currently collaborate with other facilities, institutions, etc. both nationally and internationally to develop enhancements to your data management and analysis infrastructure? Do you distribute data to others for further analysis? What metadata is collected for the purpose of data sharing?

ARM collects, integrates and then distributes derived data products from a wide variety of sources from ARM measurement sites to external contributors from NASA, NOAA, ECMWF, NWS, and others.

These derived data products are then distributed to a global user base via the ARM Archive Facility at ORNL. Metadata and provenance are collected and made available to the broader user community. This well-established workflow has been in operation for over three decades.

ARM also collaborates with its users, both individually to design new research campaigns and as projects (e.g. CSSEF, FASTER) to optimize its science delivery.

Biologists at Sandia studying carbon cycling as well as climatologists doing ARM work currently collaborate with BER, LBL, JGI (Joint Genome Institute), and the University of New Mexico. The geosciences group at Sandia is also integrating with these efforts. The University of New Mexico has received over two decades of NSF funding for its Long-Term Ecological Research Station studying nearby arid lands and also includes investigators studying polar regions.

1.2 Future Environment 2-5 years

Please describe how your scientific environment is expected to change in the coming years i.e., in terms of proposed architectures for data science infrastructures, data rates, data volumes, new analytics capabilities and new types of experiments e.g., multi-modal that require different support.

Some of ARM's future developments could be strongly influenced by the recently proposed BERAC plans for a *Biological and Environmental Research System Science Network* envisaging a stronger integration with other BER user facilities, across scales and disciplines, a wish to conduct co-located measurements across disciplines and its participation in the Biological and Environmental Knowledge Base. This would lead to an increased need for the integration of heterogeneous data, interaction with highly distributed data sources and collaboration on distributed data discovery, analysis and integration in support of result interpretation and knowledge discovery. Ideally, there would also be integration with data sets outside DOE (e.g., DoD, NASA, ECMWF, etc.) and with the geosciences.

1.2.1 Analytics Questions

What are the key future scientific questions that you expect your user communities will want to answer, how are they different from today in terms of analysis requirements (e.g., real time), what expectations will the users have of your solutions (e.g., response time, how the results would be presented, additional analytic capabilities)?

A key component of future climate science initiatives is the integration of a variety of techniques from field campaigns, to observation facilities, to laboratory experiments, and simulation/modeling across a variety of domains. An example of this integration among models, observations, and experiments across multiple scales of time, space, sciences, and system organization is illustrated below.

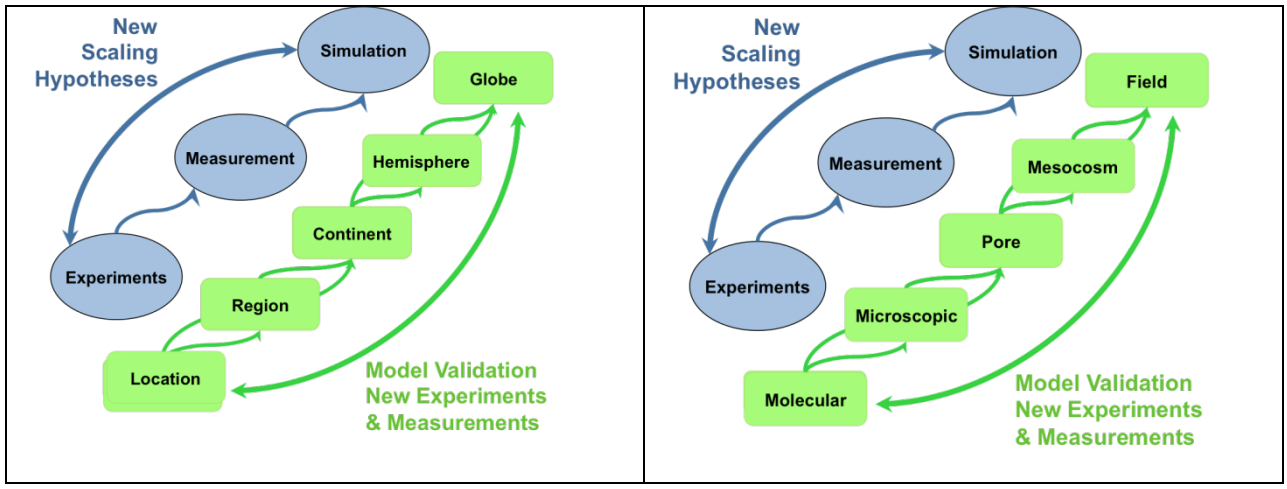

Figure 1: Integration of data across multiple scales and system organizations - Courtesy ORNL

To support this scenario, advances must be made in the analysis tools and techniques for data integration across multiple scales and system organization. Current approaches to data analysis fall short in supporting these studies.

1.2.2 Are There New Questions That Are Impossible or Impractical to Answer?

Are there questions your users would like to address, but feel it is impossible or impractical to do so with current analytic capabilities? Are there services your user community would like to see you provide, but which are impractical to offer or are currently not offered?

Many of the above-noted questions that require integration are difficult to answer given limitations in the current data infrastructure and our inabilities to distinguish correlations with causations. As a specific example, identifying the relations between molecular level "soil-omics", to plants and forest canopy, to regional climate conditions are difficult if not intractable given the state-of-the art in current technologies. Another example is the ability to request specific events from the ARM (or other BER data archives) that require higher-level constructs to be exposed and/or derived from the data. An example query might be "I would like to receive all data from a specific geographic location for days when cloud cover was in excess of 65%?".

Carbon assimilation by marine phytoplankton in the ocean is responsible for 50% of the global primary productivity (fixation of carbon dioxide, CO_2 into organic matter). Understanding how this fixation occurs and how marine environments can adapt is essential to climate science. Will ocean and terrestrial biological systems enable beneficial adaptations? Or will they exacerbate effects of climate change? Is our fundamental understanding of disparate systems in bio-, geo-, atmospheric sciences sufficient to integrate them into reliable models?

1.2.3 Major Roadblocks—Open Questions

Looking at the future environment and analytics challenges, where are the major roadblocks; please specify as much as possible current limitations and future goals (e.g., can deal with data rate of 2GB/s but not 2TB/s). You might like to use the categories of hardware, data acquisition, data management, software infrastructure, math of data analysis, and applications.

The major roadblock is the challenge of data integration given the broad diversity of data that must be brought to bear to address new questions as identified above. The overall challenge is not a data volume issue, it is a data variety issue that spans multiple data formats, different units of measure, different spatial and temporal frequencies and resolutions, and different system organization.

Other challenges involve basic system infrastructure and are identified in previous requirements analysis (The ARM ESNET Requirements analysis document).

We currently don't do a good job of integrating biological data into the climate models. Nor do we have a sufficient understanding of biological tipping points and their consequences, or of the resiliency and relevant time scales of biological contributions to climate models. The ability to begin integrating these essential factors would enable generation of more reliable models.

1.2.4 Benefit If Solved

Please describe the impact and benefits that a solution would have that would overcome specific data challenges and roadblocks listed above both in terms of facility/instrument operation and scientific impact; please prioritize those that would have in your opinion the highest impact on scientific discovery.

The primary benefit would be the ability to generate more reliable climate models and gain deeper understanding of complex systems and their emergent behaviors.

The overarching benefit to addressing these challenges would be a significant step forward in realizing a predictive modeling and understanding of the environment.

1.3 Projections of ARM Data Volume and Processing Needs

ARM

	2013 Projected need	2015 Projected need	2018 Projected need
Data Rates (GB/sec)			
Mode: burst/stream	Stream	Stream	Stream
Total volume (TBs) before reduction: per day/year	1/400	4/1600	34/12500
Total volume (TBs) after reduction: per day/year	.3/125	1.4/510	11/4000
Total volume per	1100	2300	19500

experiment (TBs) (interpreted to mean total volume across the Archive)			
Main processing mode: Central/user	Central and user	Central and user	Central and user
Total storage per year (TBs): At facility/other locations	250	1000	7500
Shared data volume per year (TBs): (transfer to other locations)	150	600	5000
Local Analysis Processing requirements by facility (range of number of cores)	100	2,000-5,000	?
Remote Analysis Processing requirements by facility (range of number of cores)	?	?	?