Data and Computational Science Challenges for Earth Science Observational Systems

D. Crichton[1], M. Little[2], G. Djorgovski[3], R. Doyle[1], Y. Gil[4], E. Law[1], J. Lemoigne-Stewart[5], C. Lukashin[6], P. Mehrotra[7], J. Nelson[8], D. Williams[9]

NASA Jet Propulsion Laboratory[1]
NASA Headquarters[2]
California Institute of Technology[3]
USC Information Sciences Institute[4]
NASA Goddard Space Flight Center[5]
NASA Langley Research Center[6]
NASA Ames Research Center[7]
USGS Eros Data Center[8]
DOE Lawrence Livermore National Laboratory[9]

1. What are the key challenges or questions for Earth System Science across the spectrum of basic research, applied research, applications, and/or operations in the coming decade?

Over the next decade, the dramatic growth of NASA's Earth Science data collections is projected to outpace the ability of scientists to analyze that data meaningfully. What has been termed the "Vs" of data (volume, variety, velocity, etc.) pose significant challenges for both Earth Science missions and researchers as traditional methods for developing science data pipelines, distributing scientific datasets, and performing effective analysis will require new approaches. The Intergovernmental Panel on Climate Change's (IPCC) Assessment Report 6, for example, predicts the growth of data to tens of petabytes. Future remote sensing projects include an increasing set of data-intensive instruments that will scale from the 100s of TBs to many several petabytes (e.g., NISAR is estimated to be in the 80 PB range) posing severe challenges to an observing system that was not designed for such a data intensive challenge. For Earth Science data archives, these massive increases demand a shift of focus from distributing whole data sets to providing online services for computation and analysis. In addition, instruments flown on NASA Earth-observing satellites will continue to generate more and more data and stress the boundaries of end-to-end data systems. These challenges taken together require new thinking in data capture, management, processing, and analysis, both onboard and on the ground, for how data systems can leverage data-driven methodologies.

Currently, the analysis of large data collections from NASA or other agencies is executed through traditional computational and data analysis approaches, which require users to bring data to their desktops and perform local data analysis.

Alternatively, data are hauled to large computational environments that provide centralized data analysis via traditional High Performance Computing (HPC). Scientific data archives, however, are not only growing massive, but are also becoming highly distributed. As a consequence, neither traditional approach provides a good solution for optimizing analysis into the future. Further, assumptions across the NASA mission and science data lifecycle, which historically assume that all data can be collected, transmitted, processed, and archived, will not scale as more capable instruments stress legacy-based systems. A new paradigm is needed in order to increase the productivity and effectiveness of scientific data analysis. This paradigm must recognize that architectural and analytical choices are interrelated, and must be carefully coordinated in any system that aims to allow efficient, interactive scientific exploration and discovery to exploit massive data collections, from point of collection (e.g., onboard) to analysis and decision support. Expanding on this point, the most effective approach to analyzing a distributed set of massive data may involve some exploration and iteration, putting a premium on the flexibility afforded by the architectural framework. The framework should enable scientist users to assemble workflows efficiently, manage the uncertainties related to data analysis and inference, and optimize deep-dive analytics to enhance scalability. These challenges are not limited to NASA. Multiple agencies are confronted with the question of how to draw scientific inference from growing, distributed archives [1].

In 2015, the NASA Advanced Information Systems Technology (AIST) Program performed a Big Data Study [2], entitled "NASA Earth Science Research in Data and Computational Science Technologies", identifying several trends as NASA continues to move towards a data-intensive agency, particularly for Earth Science missions and research. The study expanded on a road mapping effort conducted by the NASA Office of Chief Technologist that included needed technologies in Modeling, Simulation and Information Technology [3]. Critical points of the AIST study pointed to the need to look at the role of scalable data and computational technologies across the entire data lifecycle, from point of collection all the way to extracted understanding of the data. In many cases, this "data ecosystem" needs to be able to integrate multiple observing assets, ground environments, archives, and analytics, as a shift from stewardship of measurements of data to using computational methodologies to better derive insight from the data that may be fused with other sets of data. The movement to such a capability requires a careful set of steps and a roadmap for shifting the paradigm. The results of the study were shared as part of a larger workshop on Data and Computational Science in October 2015 at the IEEE Conference on Big Data [4].

Overall, we believe that NASA should consider the following as driving needs to ensure it is prepared for the next decade:

- Scale Computational and Data Infrastructures for future Earth Science mission and research needs.
- Support the application of automated data science methods, including statistical and machine learning intelligence, for deriving scientific inferences, coupled with quantifying the uncertainty of such inferences.
- Shift towards integrated data analytics where data from multiple measurements across archives can be brought together to answer specific science and applications questions and needs.
- Apply computational and data science capabilities across the data lifecycle, from flight to ground to analysis, to address increasing data demands from instruments, ground systems, and users.
- Develop capabilities that can scale appropriately to support different types of measurements from satellite to airborne to in situ.

2. Why are these challenge/questions timely to address now especially with respect to readiness?

A Data Science Working Group, first established at JPL, and expanded as part of the NASA AIST Big Data Study, identified several use cases that present challenges to future NASA missions and science research as shown in Table 1. These use cases identify capabilities that are needed in order to not just keep pace but increase the science yield from NASA Earth Science missions, instruments and data collections. They identify the need to consider new paradigms for how to construct a data intensive observing system including high performance spaceflight computing, on-board computational capabilities for event detection and data reduction, scalable communications infrastructures for both flight and ground, ground systems coupled with intelligent algorithms for feature and event detection, massively scalable computational infrastructures (e.g., high performance computing) and archives, and distributed data analytics built on data discovery techniques to support understanding of massive data collection coupled with interactive methods for visualization.

Table 1: Data Science Challenges in Different Earth Science Areas

Use Case/ Area	Description	Data Science Challenge	Enabling Mission/ Capability
Climate Modeling	Formulate hypotheses from observed empirical relationships; Simulate current and past conditions under those hypotheses using climate models; Test hypotheses by comparing simulations to observations;	Highly distributed data sources; fusion of different observations; moving computation to the	CMIP6 will move towards exascale archives requiring new approaches to evaluating models relative to observational data.

	Evaluate uncertainty of predictions originated from statistical sampling of models and observations.	data; data reduction	
Satellite Missions	Missions such as NI-SAR and SWOT will generate massive observational data. However, they are have different architectural patterns including compute intensive, data intensive, heterogeneous, etc.	Massive data rates, data movement challenges, computational scalability, archiving and distribution; onboard processing for data reduction/analysis; high-volume data transfer for ground processing	NI-SAR and SWOT require new approaches for computation, data movement, data archiving and distribution, analytics.
Applications - Hydrology (Central Valley of California)	Understanding groundwater dynamics on a regional scale using measurements from satellite, airborne and in-situ measurements. Compare against predictive models.	Distributed computation; highly distributed data sources; data fusion of multiple products; massive new satellite observations.	Integration of data from PALSAR-2, Sentinel, Grace-FO, ASO, and SMAP. Scale to support NI-SAR and SWOT. Comparison against models. Requires new architectural approaches for distributed data analytics.
Airborne Missions	Airborne missions tend to be much more agile and on-demand. They also require cost effective solutions that can fit the budget profile of airborne instruments. Integrating this into a data ecosystem provides new opportunities to quickly generate and understand various measurements and to bring airborne data into the ecosystem.	Cost effective solutions; On-demand architectures; distributed data sources; on-the-fly data processing; onboard processing for data reduction/analysis; high-volume data transfer for ground processing	Current missions such as CARVE and Airborne Snow Observatory; Future such as proposed EVI-3 and ASO follow-on missions

A major gap has been the focus on localized data analysis from instruments vs. a holistic consideration of all the observing systems and how they fit into a big data analytics view and capability. Data systems today are organized around the capture and archiving of data from specific missions or observing capabilities. However, the integration of data from multiple instruments (spaceborne, airborne, ground-based) is important for supporting scientific research, in particular, in moving from isolated data analysis to knowledge discovery through the use of a big data analytics approach. This includes making data available from multiple sources and integrating those data using intelligent algorithms and methods. In addition, as data

grow and more automated methods are in place for data discovery, this affords opportunities to improve the efficiency and effectiveness of ongoing mission operations and move it towards a data-driven approach where data are reduced onboard or at ground stations prior to archive, as well as during offline science analysis. The introduction of new approaches with interpretation of data across the lifecycle allows for informed decisions at arbitrary points in the lifecycle – allowing for mission plans to be updated, new relevant data products to be generated, etc. This same full view of the data lifecycle also will inform provenance practices, so that the relevant details including workflow – all the way back to the point of collection – can be captured to provide a basis for reproducing the end results of scientific understanding. This is a paradigm shift from how mission and science operations vs. analysis are performed today.

In looking at the challenges, Table 2 below was constructed to identify four key areas that need to be addressed to support future observational science missions, including onboard computing, ground-based systems, archives and data collection, and integrated data analytics.

Table 2: Future Data Science Capabilities for Earth Science Observation

System	2015	2025	Application to Earth Science
Onboard	Limited onboard computation including data triage and data reduction. Investments in new flight computing technologies for extreme environments.	Increase onboard autonomy and enable large-scale data triage to support more capable instruments. Support reliable onboard processing in extreme environments to enable new exploration missions.	Onboard computation for airborne missions on aircraft; new flight computing capabilities deployed for extreme environments; use of data triage and reduction for high volume instruments on satellites.
Ground Systems	Rigid data processing pipelines; limited real-time event/feature detection. Support for 500 TB missions.	Increase computational processing capabilities for mission (100x); Enable ad hoc workflows and reduction of data; Enable realtime triage, event and feature detection. Support 100 PB scale missions.	Future mission computational challenges (e.g., NI-SAR); support more agile airborne campaigns; increase automated detection for massive data streams (e.g., automated tagging of data).
Archive Systems	Support for 10 PB of archival data; limited automated event and feature detection.	Support exascale archives; automated event and feature detection. Virtually	Turn archives into knowledge-bases to improve data discovery. Leverage massively scalable virtual data storage infrastructures.

		integrated, distributed archives.	
Analytics	Limited analytics services; generally tightly coupled to DAACs; limited cross-archive, cross-agency integration; limited capabilities in data fusion; statistical uncertainty; provenance of the results	Analytics formalized as part of the mission-science lifecycle; Specialized Analytics Centers (separate from archives); Integrated data, HPC, algorithms across archives; Support for cross product data fusion; capture of statistical uncertainty; virtual missions.	Shift towards automated data analysis methods for massive data; integration of data across satellite, airborne, and ground-based sensors; systematic approaches to addressing uncertainty in scientific inferences; focus on answering specific science questions.

3. Why are space-based observations fundamental to addressing these challenges/questions?

While our purpose is not to suggest the types of measurements and their role in addressing specific science questions, it is our position that the space-based observations are a critical measurement that must fit into a broader data architecture enabled by data-driven discovery methodologies. The development of data and computational capabilities will be essential as the need to the need to capture, integrate and analyze these measurements continue to rely more and more on systematic approaches.

References

[1] National Research Council, "Frontiers in Massive Data Analysis", 2013.
[2] AIST Big Data Study, 2015. http://ieee-bigdata-earthscience.jpl.nasa.gov/references
[3] NASA OCT Roadmap, 2015. http://www.nasa.gov/offices/oct/home/roadmaps/
[4] Workshop on Data and Computational Science for Earth Science Research, IEEE Big Data Conference, 2015. http://ieee-bigdata-earthscience.jpl.nasa.gov