Challenges Facing The Geodynamics Community From The Tsunami Of Space Geodetic Observations

Susan Owen[1], Jennifer Cruz[1], Eric Fielding[1], Hook Hua[1], Paul Lundgren[1], Mark Simons[2], Paul Rosen[1], Frank Webb[1], Sang-Ho Yun[1]

Representing the Advanced Rapid Imaging and Analysis (ARIA) Project at JPL and Caltech.

1. Jet Propulsion Laboratory, Pasadena, CA
2. Seismological Laboratory, Caltech, Pasadena, CA

Introduction

The technology behind earth science observations has advanced to the stage where there are no longer just a few sources of imagery from space provided by NASA in relatively small quantities. Space agencies from around the globe are regularly launching missions to study the earth system, and data policies are becoming increasingly open. As the costs fall for installing in-situ 24-7 monitoring systems, the quantity of ground-based measurements is skyrocketing as well. While the problem we articulate here for the NSF Earth Science community is focused on the tsunami of data facing one portion of the community – the space-based geodetic community – we see the need for developing the infrastructure necessary to handle a rapidly rising flood of observational data sets as an issue facing us all. This includes the ability to process, distribute, collaborate, and understand, and preserve the current and oncoming voluminous data from Earth Observation System (EOS).

The Space Geodesy Data Tsunami

By the end of this decade, we expect nearly daily global coverage of Earth by interferometric synthetic aperture radar (InSAR)-capable satellites. This will be achieved through the upcoming launches of radar missions by the European Space Agency (Sentinel-1a and -1b), the Japanese space agency (ALOS-2), the Canadian Space Agency (Radarsat Constellation missions), the Argentine Space Agency (SAOCOM-1A/1B). NASA is planning to also launch an L-band InSAR mission in the near term. In addition, more GPS networks are installed every year, with increasing numbers capable of delivering real-time high-rate data. Observations from the future global constellation of satellites, combined with the ongoing acceleration in observations from dense GPS networks, is enabling a quantum leap in the contributions of space geodesy to studies of Earth's surface and interior. We expect to see profound impacts on the study of earthquakes and tectonics, the evolution of magmatic systems, glacier and sea ice dynamics, mechanics of landslides, and processing controlling the migration of fluids (CO_2, water, oil and gas) in the shallow crust. The near daily temporal sampling of the Earth's active deformation by these radar systems as well as sub-daily sampling from GPS, will enable scientists and engineers to develop a new generation of capabilities for rapidly responding to natural disasters (fires, earthquakes, volcanic eruptions, floods, oil spills, etc.).

The increased spatial and temporal coverage will allow for routine generation of scientifically useful data products that we have only begun to envision. For example, earthquake hazards and earthquake mechanics models of increasing fidelity would be generated from routinely produced ground surface deformation maps for all potentially active on-land faults from the integration of geodesy with conventional land-based methods. Volcano eruption forecasting and leaps forward in comparative volcanology would be enabled from the quasi-real-time monitoring of all magmatic plumbing systems by global, daily observations of all volcanoes using InSAR. Ice mechanical models that would form the basis of new predictive capabilities would be constrained for all glaciers, arctic and temperate, by tracking from space changes in their extents and velocities.

The ARIA Project

The Advanced Rapid Imaging and Analysis (ARIA) project is a joint venture co-sponsored by California Institute of Technology (Caltech) and by NASA through the Jet Propulsion Laboratory (JPL). We are currently developing a prototype end-to-end geodetic imaging data system enabling near-real-time science, assessment, response, and rapid recovery. This prototype is expected to be the foundation for an operational data processing center integrating InSAR, GPS, pixel tracking, seismology, and modeling to deliver actionable science products. Analyses of these data sets are currently handcrafted following each event and are not generated rapidly and reliably enough for response to natural disasters. The ARIA operational data processing center also plans to provide automated imaging and analysis capabilities necessary to keep up with the imminent increase in raw data from geodetic imaging missions planned for launch by NASA, as well as international space agencies.

Use Case Scenario: Volcano Science and Hazard Monitoring

Volcanoes are one area of scientific and hazard interest that will see a significant return from global short repeat InSAR coverage. For example, currently InSAR is most relevant for volcano studies as cumulative snapshots of a volcanic eruption or diking event with in-situ geodetic (tilt, GPS) providing the only sufficient temporal sampling to inform decision makers about impending eruptions. However, most volcanoes have either no in-situ geodetic monitoring or at best it is spatially aliased. InSAR, when it exists, provides important constraints on the geometric and volumetric source properties through its comprehensive spatial coverage. With the emerging constellation of satellites, and the shortening of InSAR revisit intervals, these spatially dense data sets will increasingly be able to provide insight into magma transport dynamics and become a more integrated tool for volcanic eruption monitoring.

<u>Current Scenario</u>

Recent eruptive activity at Kīlauea Volcano, Hawai‘i provides a clear example of the need for timely spatially dense observations from InSAR. On March 5, 2011, the Pu‘u ‘Ō‘ō vent on Kīlauea’s east rift zone began to subside rapidly as magma drained from beneath the cone, and a seismic swarm began about 4 km to the west. About 30 minutes later, the onset of rapid deflation was detected at the volcano’s summit. Later, the Kamoamoa fissure eruption lasted for the next ~100 hours and erupted ~2.7 million cubic meters of lava. Deformation associated with the eruption was tracked with ground-based

tilt and GPS sensors, but the fissure occurred in a sparsely instrumented area, leading to uncertainty about the likely future course of activity.

InSAR data provided a comprehensive overview of deformation associated with the eruption. ALOS PALSAR and COSMO-SkyMed (CSK) imaged Kīlauea 17 and 37 hours, respectively, after the start of the eruption, and captured subsidence of both Kīlauea's summit and Puʻu ʻŌʻō vent in addition to opening of the rift zone in the area of the eruption. Unfortunately, the advantages offered by InSAR for tracking deformation were delayed for 12 hours (CSK) to several days (ALOS): the CSK data confirmed the extent of the dike injection, but after the period of crisis had largely passed. InSAR images available within the first day would have provided important constraints on spatial extent of the dike and whether or not the fissure eruption might grow.

Future Scenario

A cloud-based data processing center such as ARIA will provide a set of monitoring data products to decision support and scientific end users. For example, volcano observatories specifying regions and products of interest via web portal or REST services interface will trigger the ARIA center to continually monitor the data archives for new relevant data, and process new scenes and GPS data as it becomes available. The system will deliver updated actionable products (e.g., interferograms, time series, GPS results, coherence maps, advisory alerts etc.) via interoperable services and data formats. All data discovery, processing, and distribution will be executed in the regional cloud endpoints closest to the data archive centers. ARIA data products will then be distributed from cloud regional endpoints closest to the intended users. Our web portal will also provide services to generate files needed to easily visualize the products via Google Earth.

In this future scenario, InSAR images with updated information on Kīlauea Volcano's deformation status would be transmitted to Hawaii Volcano Observatory (HVO) as available. By 2018, the average data latency for a repeat pass of a radar mission will be 16 hours at worst, 7 hours at best depending on data policies. The extent of a fissure eruption would be easily identified from the InSAR image well within a day of fissure eruption's initiation. This data set would provide the scientists at HVO constraints on the extent of the eruptive activity, in particular the effects of subsurface magma migration that is difficult to see from the aerial surveys. The scientists would be able to make more informed conclusions regarding the future hazard of this type of fissure eruption.

The Challenge

The potential contributions of these observation systems are vast, and only limited by our current capability to meet the computational challenges posed by them. Each of the InSAR satellites will likely produce of order terabytes per day or petabytes per year. These data are multipurpose and heterogeneous, and for the science community to fully exploit them for large-scale science problems, a much more comprehensive and scalable approach to data processing and synthesis is needed. The sheer volume of data combined with the associated processing burden and different uses for the data suggests that we cannot continue with the present model within our community whereby individual users or groups of users are expected to find a specific subset of data and learn the detailed low level processing chain for observations from each observing platform. Much of this process can be and should be streamlined and automated in such a way that it frees

scientists to focus on how to further exploit the observations in research and applications. This will require developing a range of data and processing infrastructure capabilities to serve the needs of both the individual scientist and centralized analysis centers.

With the onslaught of "data tsunami"—particularly from high temporal and spatial resolution geodetic observations—an ARIA-like cyberinfrastructure supporting just the high-volume and low-latency processing alone is insufficient. We as an Earth Science community must also support multi-faceted capabilities including effective federated data handling, data discovery, data and service interoperability, as well as collaborative methods to understand and share scientific results among cross-disciplinary communities.

Our team is developing an analysis center for automated data processing of InSAR and GPS data that will enable detecting precursory deformation as well as generation of multi-level data products for science and monitoring of on-going hazardous activities around the globe. As our contribution to the EarthCube discussion of science requirements, we discuss here our current approaches for solutions and the science and design requirements as we see them for the addressing the "data tsunami" problem facing space geodetic community.

A Two-Tiered Cloud-Based Computing Approach

The vision for cloud-based computing for collaborative science extends the user's computer workstation and local processing capability to a vast infrastructure on the cloud. Here we present two approaches: one where the user is an individual scientist analyzing SAR data, and another where the cloud-based computing user is an automated data processing system for geodetic imaging. We see both solutions as necessary since science advances by scientists that can customize the analysis in new and novel ways and by scientists who have easy access to massive data sets for analysis or combination with complementary data in integrative models. In addition, the needs of the hazard community must be met with cloud-based automated data processing in order to ensure reliable data availability.

Numerous analyses currently suffer from drastic delays in their pipeline simply because the processing is confined to locally available infrastructure. In contrast, a community solution would exercise the large number of independent processors at cloud sites to process much of the data in parallel, so that simple temporal reductions such as stacking occur very quickly. For more complex analyses, one of the primary goals of this approach is to be able to organize the computing in a manner that facilitates quick throughput. This would extend far beyond simple code parallelization by re-architecting typical flow pipelines to take advantage of the virtually limitless compute capacity available concurrently in the cloud environment.

"Individual Scientist" User for SAR analysis

This approach is designed to allow scientists and engineers to produce radar interferograms and higher-level products without requiring local computational or data storage resources. Rather, the user can select a set of data to be processed, the algorithms with which to analyze the data, and receive the relevant processed subset on a mobile device. One could build enabling technologies that allow, for example, scientists in the

field to access InSAR analyses, limited to only cellular network connectivity, and to specify how the data will be processed and what products will be produced.

SAR data are archived at several places around the world, for example, at the Alaska Satellite Facility (ASF) or the UNAVCO WInSAR archive in the U.S., at ESA archive sites in Europe, and at JAXA in Japan. Data are also available commercially from many providers at cost to the users. Our solution would find the location of data in one or more of these archives for a particular study locale, and move the desired scenes to a cloud-based processing facility. The selected processing algorithm would then reduce the raw data to high-level data products. Finally, the product in a useful format could be downloaded to the users' device. A goal would be to minimize end-product latency to minute to hour retrieval times for the products after origination of the order, especially for large data sets coming from new instruments. While individual scenes available on archives such as ASF or WInSAR can be processed in the best case in comparable timeframes, large data sets cannot. A community solution would establish the framework for scalability to allow hour-scale processing of large data sets.

"Data System" User for Geodetic Imaging

This approach is design to provide an integrated service for InSAR and GPS data processing in an elastic computing paradigm that is ideal for monitoring and response to globally distributed hazards. In addition to more informed decisions by monitoring agencies, automatically and centrally provided products would enable greater understanding of processes leading up to, during, and after natural and man-made disasters. The global coverage offered by satellite-based SAR missions and rapidly expanding GPS networks can provide orders of magnitude more observations if there is a centralized processing system that can efficiently analyze the voluminous data, and provide users the tools to access data products for their regions of interest.

Unlike other cloud approaches that are algorithm-specific, the ARIA team aims to cloud-enable generic science data system components so that arbitrary execution code is run "embarrassingly parallel" on the cloud. Our current prototype ARIA science data system contains science data system processing architectural elements of workflow management, resource management, job management, data repository, and data discovery. We plan to augment these components to interface with cloud services.

Data Product Requirements

The following is a high level list of requirements for data products that are produced by any decentralized processing infrastructure or by a centralized automated center. The customers here are both the science and hazard communities.

1. Rapid, reliable access to data, products. Timeliness and consistency is critical for monitoring agencies, but is also enables scientists to focus on higher level analysis of data sets if they can rely on a source of processed images or time series.
2. Multiple levels of products made available for multiple levels of users from expert to non-scientists. Geophysicists may need a regional scale 3-D surface deformation map of an earthquake event, while responding government agencies want a high-level damage assessment map of a particular urban area.

3. Interoperable metadata. All data and derived products should be accompanied by complete metadata to enable correct use of them, and be reproducible by other scientists. The metadata should be available in common standard format for interoperability among different tools and institutions. For example, the series of ISO 19115-related metadata model for geographic metadata increasingly being used across the Earth science communities, while ISO 19139 is being used as the interoperable encoding for representation. When common standard format is not available, format-converting tools should be provided.
4. Easy federated discovery and search of data products, including option for data products to be pushed when available. Given the large volume of distributed data products that will be available, along with their decadal-scale temporal coverage and global coverage, it will be necessary to develop methods for scientists and non-scientists to easily discover and access these distributed data holdings.
5. Sophisticated tools for learning about and exploring data sets and data products. Usage instructions for data sets and derived products should be provided for multiple levels of users, sophisticated tools for exploring the products should be provided if the usage is not straightforward with publicly available standard tools.

Design Requirements

The following is a high-level list of requirements for the design of any decentralized processing infrastructure or centralized automated data system. The customers here are both the science and hazard communities.

1. Loosely-coupled architecture to enable rapid assimilation of new technology advancements in different components as they develop at different rates.
2. Access to elastic compute resources to ensure that the system scales to changing processing demand (e.g., in the wake of a natural disaster and the need to process the associated data)
3. Distributed data storage to facilitate low-latency data access across geographically dispersed compute nodes.
4. Interoperable metadata models and encoding formats that are infused into tools, data systems, and used across different communities.
5. Generation of higher-order data products from the observational data enabling greater understanding by cross-disciplinary communities. To service the hazards and disasters communities, "high-ordered" actionable data products can yield greater impact to societal needs.
6. Federated data discovery and access enabling scalable handling of "big data". The expected tsunami of InSAR data alone cannot be effectively handled by one data center.
7. Data product preservation and stewardship enabling product provenance, transparency, and reproducibility.
8. Visualization environments with sufficient network and processing bandwidth to enable innovation and discoveries not otherwise possible.

We look forward to continued discussions with EarthCube members on what cyber infrastructure is needed by the Earth Science community to face this "data tsunami" challenge.