NASA Science Mission Directorate
Research Opportunities in Space and Earth Sciences – 2017
NNH17ZDA001N-ACCESS
A.42 Advancing Collaborative Connections for Earth System Science

The Advancing Collaborative Connections for Earth System Science (ACCESS) Program solicits projects to develop and implement technologies to effectively manage, discover and utilize NASA's archive of Earth observations for scientific research and applications in support of NASA's Earth science research goals. The 2017 ACCESS announcement specifically sought technology developments for one or a combination of the following areas: Machine Learning, Advanced Search Capabilities and Cloud Optimized Preprocessing and Data Transformation technologies to improve data discovery, usability, and management of large and complex Earth science data sets from EOSDIS.

A total of 39 proposals were received for this announcement. All proposals were peer evaluated using an expert panel review. The Earth Science Division of NASA's Science Mission Directorate selected five proposals for two-year awards pending satisfactory budget and work plan negotiations. The total first-year funding for these proposals is approximately 3.2 million dollars. These projects will improve and expand the use of NASA's Earth science data by leveraging modern techniques to address challenges in the management and analysis of large and complex Earth science data sets

---

Anthony Arendt/Applied Physics Laboratory, University of Washington, Seattle
Community tools for analysis of NASA Earth Observing System Data in the Cloud
17-ACCESS17-0003

Data intensive scientific workflows are at a pivotal time in which traditional local computing resources are no longer capable of meeting the storage or computing demands of scientists. In the Earth System Sciences (ESS) community, we are facing an explosion of data volumes where new datasets, sourced from models, in-situ observations, and remote sensing platforms, are being made available at prohibitively large volumes to store at even medium to large High Performance Computing (HPC) centers. NASA has estimated that by 2025, it will be storing upwards of 250 Petabytes (PB) of its data using commercial cloud services (e.g. Amazon Web Services [AWS]). Availability of these data in cloud environments, co-located with a wide range of computing resources, will revolutionize how scientists use these datasets and provide opportunities for important scientific advancements. Fully leveraging these opportunities will require new approaches in the way the ESS community handles data access, processing and analysis. These technologies will be deployable on commercial cloud infrastructure where Earth Observing System Data and Information System (EOSDIS) is anticipated to be stored. At present, tools for working with these datasets consist of convenient interfaces for discovering and downloading data (e.g. NASA's Earthdata search) from individual Distributed Active Archive Centers (DAACs). We anticipate that the transition to cloud storage for many of these DAACs will bring immense opportunities and specific challenges to researchers. Our proposal will facilitate the ESS community's transition into

cloud computing by developing technologies that build on existing open-source tools (e.g. Python, Jupyter) by integrating building on top of the growing Pangeo ecosystem. Our first task will be to deploy a scalable cloud-based JupyterHub on AWS for community use. JupyterHub is a multi-user, multi-language interactive computing environment that facilitates open-ended, exploratory analysis and data visualization. Content ('notebooks') developed on JupyterHub are both functional and fluid; in the manner of an 'executable paper' combining data, processing and interpretation, a necessary departure from traditional publication as a sequence of static artifacts.
Our second task will be to integrate existing NASA data discovery tools with cloud based data access protocols. While existing data discovery tools, such as CMR/GIBs, provide convenient access to dataset metadata but navigating the access, retrieval, and processing steps for these datasets is left to individual users. We will develop an advanced Python API that leverages high-level tools like Xarray and Dask allowing scientists to accelerate their analysis. Integration of this API with the Pangeo ecosystem will provide our API with cutting edge scientific tools for pre-processing, regridding, machine learning, and visualization.
Our third task will leverage our advanced API for data discovery and processing to provide an advanced, cloud-optimized framework for remote data retrieval. Our approach to a data retrieval system goes beyond simple slice and download operations (e.g. OpeNDAP) and leverages our advanced API for data discovery, access, and processing to also provide server-side perfunctory processing.
We will demonstrate the use of these tools with several datasets including North American Land Data Assimilation System (NLDAS), Gravity Recovery and Climate Experiment (GRACE), and Sentinel-1 synthetic aperture radar. The example applications will serve as templates for the broader community and real-world applications for evaluation of the cloud services and applications we develop. We also propose to help accelerate a shift in the ESS culture toward cloud computing by providing short but intensive training opportunities. Our work will provide new ways for scientists to collaborate and make full use of NASA satellite datasets.

Patrick Heimbach/University of Texas Austin
Data Access and the ECCO Ocean and Ice State Estimate
17-ACCESS17-0028

The Estimating the Ocean Circulation and Climate (ECCO) global ocean state estimation system is the premier tool for synthesizing NASA's diverse Earth system observations into a complete physical description of Earth's time-evolving full-depth ocean and sea ice system. ECCO state estimates are of particular significance to NASA because on their own, all satellite observations, although global in coverage, remain sparse in space and time relative to the inherent scales of ocean variability, and are blind to the ocean's interior. With increased streams of data and better spatial resolution the scientific utility of the product is increasingly limited by (1) the inability to automate observing network ingestion and update in a rapid, robust manner, (2) the lack of tools for embedding the state estimate into NASA's Earth Observing System Data and Information System

(EOSDIS) framework, and (3) the lack of capabilities to perform efficient online data analysis.

To overcome these hurdles we will develop and implement a production-ready cloud-native storage and data analysis system, called ECCO-Cloud, to manage the preprocessing and transformation of NASA Earth Science data and data products The following three high-level goals will be accomplished:

I) expand and accelerate in a sustainable and scalable manner the integration of NASA Earth system data into ECCO through automated preprocessing and transformation;

II) radically streamline the integration of updated ECCO products into NASA's Earth Observing System Data and Information System (EOSDIS), specifically the Physical Oceanography Distributed Active Archive Center (PO.DAAC)

III) facilitate and expand the scientific utilization of NASA remote sensing data integrated in ECCO by the growing community of interdisciplinary researchers in the oceanographic, sea-ice, sea level rise, and climate science fields.

The proposed work leverages (1) the latest ECCO ocean global state estimate, (2) new software tools developed to display, analyze, extract, subset, reproject, and download ocean physical parameters from the ECCO state estimate [temperature, salinity, currents, atmosphere-ocean heat fluxes, sea level, etc.], (3) experience in hosting and rapidly accessing the tens of gigabytes of binary output files that comprise the complete ECCO state estimate, and (4) new developments that now allow new simulations based on ECCO's OGCM to be run on the Amazon Elastic Compute Cloud (Amazon EC2).

Recently, progress has been made in the development of software tools that allow rapid online access (through pre-caching) and analysis of a subset of ECCO output via interactive web pages. These software tools were development for NASA's "Sea Level Portal" (see https://sealevel.nasa.gov/data-analysis-tool/). In the proposed project we intend to leverage and significantly extend these tools so that a much larger set, if not all, ECCO ocean parameters are made similarly available. The tools to access and conduct interactive analyses of these ocean parameters will be implemented on the ECCO's new website. Users will be provided links to the appropriate EOSDIS Distributed Active Archive Centers (DAACs) such as PO.DAAC and NSIDC DAAC to access the original ocean and ice data products used in ECCO. This will also enable NASA's common access interface through the Earthdata enterprise Common Metadata Repository (CMR) and the Earthdata Search (https://search.earthdata.nasa.gov) capability.

Finally, now that ECCO's OGCM has been ported to run on Amazon EC2, we will provide an online front-end for users so that they can (a) reproduce the full state estimate, and (b) formulate and conduct their own experimental simulations and then seamlessly analyze the output of those simulations using the same data analysis tools.

We expect that the proposed work will contribute to fuller utilization of NASA Earth System data - especially ocean data and ECCO products - by the research community.

Hook Hua/Jet Propulsion Laboratory
Multi-Temporal Anomaly Detection for SAR Earth Observations

17-ACCESS17-0026

Synthetic aperture radar (SAR)-based geodetic imaging has revolutionized Earth science research in disciplines such as solid earth, ecosystems, and cryosphere. Yet the ability to effectively utilize SAR data for research, long-term monitoring of spatial areas of interests (AOIs), and rapid hazard response has been limited due to processing complexity, data volume sizes, and latencies in the end-to-end process. For example, barriers in urgent response include the lack of automated data triggers from forecasts, the need for specialized processing parameters that currently rely on expert intervention, and the manual delivery of actionable science data products to the decision support communities. Our team's AIST-2011 and AIST-2014 efforts towards an Advanced Rapid Imaging and Analysis (ARIA) data system has demonstrated the capability to automate high-volume SAR image analysis in a cloud computing environment, but a limiting factor has been the continued need for expert analysis for detection of features in the L2 data products as well as transients in the L3 time series data products. Decision support products, which often are most useful if they are generated rapidly and with simplified information (e.g.,damaged or not damaged or flooded or not flooded), often require change detection-based approaches utilizing the before & after event scenes to be processed often with threshold values set on the underlying SAR measurement values of either amplitude or coherence. These steps requiring a human-in-the-loop have become a bottleneck for rapid & reliable exploitation of geodetic data for both long-term monitoring and event response with SAR data. If done correctly, SAR data can also be used to monitor for events such as volcano inflation or precursory signals to landslide events. Automating these time domain-based feature detection have even larger barriers due to processing of large temporal co-registered data stacks, processing complexity, as well as human expertise needed to assess the time domain signals.
Machine learning approaches for earth science data have typically been applied on these types of single scene feature detection. In our AIST-2014, we prototyped automated classification of phase unwrapping features in processed L2 inteferograms from the Sentinel-1A/B data streams. For change detection techniques however, most methods have focused on paired "before/after" observations. Now with more availability of low-latency and global multi-temporal remote sensing data, opportunities exists to exploit detection of more time dependent features of highly temporal earth science observations. Unlike the current state of the art pair-wise change detection techniques, multi-temporal spatial prediction techniques that leverage long-term historical observations yield more accurate and more interpretable predictions.

We will: (1) on-ramp onto the cloud computing platform, an automated and large-scale pre- processing of multi-temporal data stacks, and (2) develop automated large-scale machine learning analysis of multi-temporal transient detection and precursory signal analysis of multi-temporal stack datasets. We will demonstrate real science value by exploiting NASA's Sentinel-1A/B archive in preparation for the upcoming NISAR mission. All input data will be acquired by discovering and accessing ESDIS DAAC archive data holdings via queries to the ESDIS Common Metadata Repository (CMR). Our demonstration will pre-process co-registered stack of single looks complex (SLC) data in the AWS cloud. This is both computationally intensive as well as disk I/O

intensive and therefore fits the need for cloud optimized preprocessing. The generic time-dependent anomaly detection approach will exercise example use cases such as automated landslide detection, automated volcanic uplift early detection, and/or automated detection of pre-event time-series patterns vs co-event flood detection.

Michael Rilee/Rilee Systems Technologies LLC
STARE: SpatioTemporal Adaptive-Resolution Encoding to Unify Diverse Earth Science Data for Integrative Analysis
17-ACCESS17-0039

OBJECTIVES: With "STARE: SpatioTemporal Adaptive-Resolution Encoding" we address Focus 2.1.3 "Cloud Optimized Preprocessing and Data Transformation." Current Earth Science data processing features large, centralized archives providing exceptional browse and search capabilities used by researchers who identify then download data in file form to local compute/storage resources for preprocessing and integration prior to analysis. This data flow forces end-users to devote scarce resources to support the transfer, storage, and management of archived data, as well as specialist expertise in the various different kinds of data sets of their research domain. We propose to simplify this flow by moving preprocessing activities to the archived data, eliminating the costs of transferring, creating, and maintaining redundant, idiosyncratic local archives, developed by researchers who are generally not archivists, nor the expert producers of the original data. With STARE providing a unifying platform for diverse data models (swath, point, grid), EOSDIS data archives will be able to produce higher-level products made to order for end-user researchers.

METHODS: The critical, new technology we apply is STARE. STARE's spatial component (SC) has descended from the Hierarchical Triangular Mesh (HTM) spherical indexing originally developed for the Sloan Digital Sky Survey, in which storage and computational efficiency was key. The STARE/SC recursively divides the Earth's surface into a set of quad-trees allowing any point on Earth to be identified with a single number. The STARE temporal component (TC) has similar properties. For observations, these STARE indices contain both location and resolution information, promoting efficient data placement on distributed, cloud resources minimizing costly data transport between nodes for operations such as joining, intersecting, (conditional) subsetting, and re-gridding diverse datasets. STARE automatically co-aligns diverse data in the cloud, placing spatiotemporally close data on the same compute/storage node, for a relatively small cost in metadata.

STARE thus allows diverse data to be efficiently integrated for analysis in the cloud, providing a foundation on which existing tools and processing methods can be placed. Much capability, e.g. preprocessing, searching, visualization, etc., has been developed to support researchers' use of Earth Science data. In the course of the proposed work, we will show how existing tools and methods benefit from the STARE-enabled platform. This can be via a tight integration as has been done, for example, incorporating STARE with the distributed array database SciDB, along with re-gridding functions, and fast

parallel, geographic intersections. Or current tools may simply be applied to the results of STARE-enabled distributed processing, e.g. fast granule intersection, in a more conventional, but cloud-based, data processing flow.

SIGNIFICANCE: As a unifying platform, STARE supports conventional processing, analysis, and visualization tools, bringing the opportunity for massively increasing the amount of data researchers can use. In the longer term, as tools evolve to take greater advantage of STARE's integrative capabilities, we can move from the current focus on the expensive low-level manipulation of data files to an ability to interact with Earth Science data at a higher level, with query-based declarative tools and user interfaces that favor scientific inquiry rather than data management. At the very least, STARE helps automate critical spatiotemporal functions while making efficient use of cloud computing, promising to eliminate the need for researchers to devote time, money, and expertise to the redundant transfer of archived data to their own, local systems. The time and effort saved improves scientific quality and the productivity of the current researchers and reduces the cost-of-entry for others who might seek value from EOSDIS data resources.

Jeff Walter/NASA Langley Research Center
Systematic Data Transformation to Enable Web Coverage Services (WCS) and ArcGIS Image Services within ESDIS Cumulus Cloud
17-ACCESS17-0014

The NASA Earth Science Data and Information System (ESDIS) Project has recently launched the Cumulus prototype which provides a scalable cloud-based platform to ingest, archive, distribute, and manage Earth Science data within the Amazon cloud environment ("Earth Science Data," 2017). One of the goals for the Cumulus project is to increase the flexibility and effectiveness of NASA Earth Science data access and distribution to end users. Earth science data is geospatial in nature. However, according to the Geospatial Data Abstraction Library (GDAL) Enhancements for ESDIS (GEE) Assessment ("GDAL Enhancements," 2016), many Earth science data products (e.g. MOPITT, CERES, to name a few) are difficult to access and use within commercial off-the-shelf (COTS) and open source Geographic Information Systems (GIS) software, such as Esri's ArcGIS and the open source QGIS software.

According to the 2017 American Customer Satisfaction Index (ACSI) survey, ArcGIS is the most used software tool/package at 64%, followed by QGIS (37%), ENVI (32%) and Excel (27%) to work with NASA Earth science data ("ACSI Reports," 2018). We propose to develop geospatial data transformation plugins that could be used within the ESDIS Cumulus environment to serve out transformed MOPITT and CERES data product(s) as OGC Web Coverage Services (WCS) and Esri ArcGIS Image Services. These services will then be easily consumed into COTS GIS software such as ArcGIS and QGIS. These plugins will perform the transformation to fix the data issues (e.g. incorrect image sizes, orientation, multidimensional variable interpretation, georeferenced metadata recognition, etc.) and this extensible framework can be further

expanded to add new plugins as additional issues are identified and addressed. In addition, ArcGIS Image Services and Web Coverage Services provide easy access to actual data values, not just static images, and alleviate the need to navigate complex and storage-intensive processes of raw data downloads. The importance of delivering geospatially enabled web services is emphasized through initiatives like the establishment of the NASA Earth Science Data System (ESDS) Geospatial Web Services Working Group (GWSWG), the NASA Big Earth Data Initiative (BEDI), and the ESDIS Cumulus Seamless 360 Degrees of Services that includes the implementation of a common user-facing API such as WCS. By developing these geospatial data transformation plugins and subsequently providing the corrected data as OGC and Esri web services, it will remove the barrier of transforming each NASA data product after download since it will be correctly served, leading to more easily accessible NASA data products by the Earth Science community.

This proposal will utilize Amazon Web Services (AWS) Step Functions and Lambda Functions as also utilized by the ESDIS Cumulus prototype, though if the ESDIS Cumulus system is not operationalized, our project can be decoupled and operate independently within a commercial cloud environment. We will approach this problem in three phases. Phase one, planning, will focus on the AWS workflow structure. Phase two, development and implementation, will deploy the AWS Step Functions and Lambda Functions used to orchestrate a workflow of customized micro-services executing GDAL transformations in order to geospatially enable and serve a new cloud-optimized MetaRaster Format (MRF) product as an OGC Web Coverage Service and ArcGIS Image Service. Phase three, testing, will allow us to measure the discoverability of NASA Earth Science data web services within major data catalogs (e.g. Earthdata Search, GeoPlatform, ASDC ArcGIS Portal, etc.), and their use, speed, accessibility and analytical capabilities within QGIS, ArcGIS, custom web mapping applications, and Data Cubes.