CDI FY17 Request for Proposals

Improving Science with Data Science: USGS Icefields-to-Oceans

Submission Title: Improving Science with Data Science: USGS Icefields-to-Oceans

Lead PI: Shad O'Neel

Mission Area: Climate and Land-Use Change

Region: Alaska

Organization: Alaska Science Center

OrcId: 0000-0002-9185-0144

Phone: 9077867088

Email: soneel@usgs.gov

City: Anchorage

State: AK

Co-PIs and Collaborators:

Science Support Framework Element 1: Data Management

Science Support Framework Element 2: Science Data Lifecycle - Preservation

Science Support Framework Element 3: Communities of Practice

In-Kind Match: $49,000.00

List of anticipated deliverables from the project: Our deliverables will include automation routines (scripts), a data release, and a white paper describing the workflow developed through the effort.

Lead Cost Center: Alaska Science Center

Notes, Comments:

Project Description: We propose advances and efficiencies in cloud-based data management including automation routines that can be broadly applied among USGS researchers.

Total Budget: $41,191.00

Improving Science with Data Science: USGS Icefields-to-Oceans
PI: Shad O'Neel- Alaska Science Center

Summary
Interdisciplinary Earth science projects enable scientific insight and cost savings, especially when data management fosters transparency and reproducibility. These projects produce 'big data', both in terms of volume and diversity, however, our computational data science infrastructure has not kept pace. Newly established transparency and reproducibility policies require programmatic investment to ensure maximum value. A need exists to develop novel ways of sharing, hosting and preserving our data to maximize its scientific impact while simultaneously minimizing formatting and preservation efforts of science staff. This proposal uses data from a recent interdisciplinary data collection effort to focus on a flexible and accessible cloud hosting and preservation approach for diverse and integrated data sets. Our work will focus on two main tasks: streamlined metadata production and compliance with USGS Fundamental Science Practice and Information Technology policies.

During 2016, the USGS glaciers and climate group, along with academic and agency collaborators, assembled a unique data set from Wolverine Glacier Alaska to better characterize glacier-ecosystem interactions. Data sets span physical, chemical and biological disciplines (field, remote sensing and model data), and are geospatially integrated. Our science is aligned with several of the 2016 USGS priorities including: water resource security, advancing remote sensing and mapping, engaging the world with the USGS mission and planning and developing our workforce. Our early efforts resulted in a university and cloud-hosted database and API (http://www2.gi.alaska.edu/~jrich/web_pages/main.html; http://aaarendt.github.io/ice2ocean/), which houses cryospheric data pertaining to the Alaska Region. However, this existing framework needs focused attention and streamlining to produce integrated products from integrated data collection. All USGS efforts will be coordinated with Univ. Washington collaborator Anthony Arendt and his data science team to maximize efficiency and prevent redundant efforts. Our requirements and goals are not unique to our project, and the work proposed here will be generalized to apply broadly across USGS data science needs. Our approach will be to:
1. Ensure cloud-access compliance for all team members (federal and academic).
2. Develop automated database formatting and ingestion methods
3. Streamline FGDC compliant metadata production for subset extractions from the database.

By standardizing and automating workflow early in the data management lifecycle, we focus our work on the "how" aspects of the CDI science Support Framework. In particular, working to achieve a streamlined, flexible approach to data integration and preservation that can be disseminated to a broader community of practice while complying with FSP, IT and OSQI requirements. This investment will thereby enhance both science and data science aspects of USGS research. Our initial efforts into the cloud-based data hosting yielded unexpected discovery as demonstrated in a 2015 publication directly cited by President Obama during his visit to Alaska (Larsen et al., 2015). We expect this work will yield substantial gains in the data release process, now required for USGS researchers prior to publishing their work. Efficient data release processes have manifold benefits, ranging from increased credit for data collection to more, transparent and reproducible science.
If funded, our efforts for this proposal will be structured such that all programming is generalized, written in a way to extend beyond our immediate needs. We will provide documentation in the form of a white paper or USGS publication detailing our approach towards transparent, reproducible, and efficient integrated data management.

Larsen, C.F., E. Burgess, A. Arendt, S. O'Neel, A. Johnson, and C. Kienholz. 2015. Surface melt dominates Alaska glacier mass balance. Geophys. Res. Lttrs. 42. Doi: 10.1002/2015GL064349.

Budget

Budget Category	Federal funding Requested	Matching funds proposed
1. Personnel (Salaries, including benefits)		
Federal Personnel Total:	$25000	$40000
Contract/Collaborator Personnel Total:	$0	$0
Total Salaries:	$25000	$40000
2. Travel Expenses		
Travel total	$2000	$2000
Other expenses	$0	$0
Total Travel	$2000	$2000
3. Other Direct Costs		
Equipment	$7000	
Publications		$2000
Office Supplies, Training		$5000
Total Other Direct Costs	$7000	$7000
Total Direct Costs	$34000	$49000
Indirect Costs (21.15%)	$7191	
Grand total	$41191	$49000

Justification: Funds are being requested for partial support of GS-9 data science position in the ASC glaciology group. Salary funds are matched in multiple ways including funds from the USGS Alaska Region and collaborator salary at University of Washington. The position has been filled and work begins November 2016. Funds are requested for 2 people to travel to CDI for an event. Funds are requested for cloud hosting of the Ice2O database, as well as storage and processing hardware.