CDI FY17 Request for Proposals

Exploring the USGS Science Data Life Cycle in the Cloud

Submission Title: Exploring the USGS Science Data Life Cycle in the Cloud

Lead PI: Nadine Golden

Mission Area: Natural Hazards

Region: Pacific Region

Organization: Pacific Coastal and Marine Science Center

OrcId: orcid.org/0000-0001-6007-6486

Phone: 8314607530

Email: ngolden@usgs.gov

City: Santa Cruz

State: CA

Co-PIs and Collaborators:

Type: CO-PI

Name: Joseph Hughes

Mission Area: Water

Region: Headquarters

Organization: Office of Ground Water

OrcId: 0000-0003-1311-2354

Phone: 7036485805

Email: jdhughes@usgs.gov

City: Tampa

State: FL

Type: CO-PI

Name: Andrew Stevens

Mission Area: Natural Hazards

Region: Pacific Region

Organization: Pacific Coastal and Marine Science Center

OrcId: 0000-0003-2334-129X

Phone: 8314607424

Email: astevens@usgs.gov

City: Santa Cruz

State: CA

Type: Collaborator

Name: Jeremy White

Mission Area: Water

Region: Southeast Region

Organization: Tampa Office, FLWSC

OrcId: 0000-0002-4950-1469

Phone: 5129273585

Email: jwhite@usgs.gov

City: Tampa

State: FL

Type: Collaborator

Name: Richard Signell

Mission Area: Natural Hazards

Region: Northeast Region

Organization: Woods Hole Coastal and Marine Science Center

OrcId: 0000-0003-0682-9613

Phone: 5084572229

Email: rsignell@usgs.gov

City: Woods Hole

State: MA

Science Support Framework Element 1: Science Data Lifecycle - Analysis

Science Support Framework Element 2: Science Data Lifecycle - Processing

Science Support Framework Element 3: Science Data Lifecycle - Publishing/Sharing

In-Kind Match: $40,000.00

List of anticipated deliverables from the project: We will create an implementation of the THREDDS Data Server and JupyterHub in the USGS Cloud Hosting Solutions environment, and all components of the Data Life Cycle will be tested in the context of specific USGS CMG projects. Specifically, we will spin up a Microsoft Windows server instance to run the Delft Flexible Mesh model, so data will be created directly in the Cloud. We will also spin up multiple Linux server instances and deploy JupyterHub and the THREDDS Data Server via Docker containers. JupyterHub will allow Python, R and Matlab workflows to run in the Cloud, next to the data, while the researcher simply interacts using a modern web browser. We will also provide enhancements to the existing pyugrid Python package that implements the UGRID community conventions for unstructured grid models.

Lead Cost Center: Santa Cruz

Notes, Comments:

Project Description: With the advent of the USGS Cloud Hosting Solutions, web technologies like the Jupyter project and web services for data distribution like the THREDDS Data Server, we should now be able to conduct most of the Science Data Life Cycle components in the Cloud, where we can share and scale our hardware resources, share software environments, perform analysis close to the data, and preserve, publish and share our data for public use. We propose here to implement the THREDDS Data Server and JupyterHub in the USGS Cloud Hosting Solutions environment, and test all the components of the Data Life Cycle in the context of specific USGS CMG projects. We will also work to enhance the python tools used in this environment in collaboration with the USGS Water Mission Area.

Total Budget: $27,000.00

SECTION 1. PROJECT SUMMARY

Project Title: Exploring the USGS Science Data Life Cycle in the Cloud

Name of USGS Lead Principal Investigator: Nadine Golden

Project Summary:

The USGS Science Data Life Cycle has been characterized by CDI as "Plan, Acquire, Process, Analyze, Preserve, Publish/Share". Traditionally in the USGS, data is processed and analyzed on local researcher computers, then moved to centralized, remote computers for preservation and publishing (ScienceBase, Pubs Warehouse). This approach requires each researcher to have the necessary hardware and software for processing and analysis, and also to bring all external data required for the workflow over the internet to their local computer. With the advent of the USGS Cloud Hosting Solutions, web technologies like the Jupyter project and web services for data distribution like the THREDDS Data Server, we should now be able to conduct most of the Science Data Life Cycle components in the Cloud, where we can share and scale our hardware resources, share software environments, perform analysis close to the data, and preserve, publish and share our data for public use.

We propose here to implement the THREDDS Data Server and JupyterHub in the USGS Cloud Hosting Solutions environment, and test all the components of the Data Life Cycle in the context of a specific USGS CMG project: Modeling climate impacts on flooding in Puget Sound. This project is a particularly useful prototype for exploring in the Cloud because the "data" is computer generated and requires substantial resources for analysis. The model data also cannot be effectively distributed on ScienceBase because it is too large (TB), and ScienceBase does not yet support OPeNDAP, the de facto community standard for allowing extraction of array based model data over the internet. It is also useful because the model data produced by the Delft Flexible Mesh hydrodynamic model used on this project is on an unstructured mesh that uses both triangular and rectangular elements, and as part of this project we will enhance the existing *pyugrid* Python package that implements the UGRID community conventions for unstructured grid models. We will work with collaborators Joseph Hughes and Jeremy White from the Water Mission Area to ensure that *pyugrid* can be used with the next release of MODFLOW (MODFLOW 6) groundwater model which will have native support for unstructured grids as well. This will enable them to take advantage of *pyugrid* in their flopy Python package for MODFLOW, instead of reinventing the wheel, and will enable a demonstration of interoperability for the *pyugrid* package. This work on *pyugrid* will be done by Axiom Data Science, who have been involved with the development of both *pyugrid* and the Python module for converting the custom binary MODFLOW output to standardized NetCDF.

Specifically, we will spin up a Microsoft Windows server instance to run the Delft Flexible Mesh model, so data will be created directly in the Cloud. We will also spin up multiple Linux server instances and deploy JupyterHub and the THREDDS Data Server via Docker containers. JupyterHub will allow Python, R and Matlab workflows to run in the Cloud, next to the data, while the researcher simply interacts using a modern web browser. Nothing is installed researchers local computer. Both the original data and data products will then be served via the THREDDS Data Server, which provides OPeNDAP services that allows public users to extract just the data they need from the simulations. The THREDDS Data Server also produced ISO metadata records that can be fed into ScienceBase and receive a DOI, thus allowing an approved and effective method for publishing model output. The code development on *pyugrid* will take place on GitHub, ensuring openness and longevity after the CDI-funded project is over.

Although a specific workflow and specific types of model data will be tested, the results of this CDI project will be useful to all USGS modelers, as this represents a general approach to performing, analyzing and publishing data in the Cloud, making the science conducted by our researchers more effective and efficient. We will welcome contributions and participation from other modeling groups in the USGS.

SECTION 2. ESTIMATED BUGET

Budget Category	**Federal Funding “Requested”**	**Matching Funds “Proposed”**
1. PERSONNEL (SALARIES including benefits):		
Federal Personnel Total:		$40,000
Contract/Collaborator Personnel Total:	$20,000	
Total Salaries:	$20,000	$40,000
2. TRAVEL EXPENSES:		
Travel Total (Per Diem, Airfare, Mileage/Shuttle) x # of Trips:	$3,000 (one trip; PI to CDI Annual Meeting in May 2017)	
Other Expense (e.g. Registration fees):		
Total Travel Expenses:	$3,000	
3. OTHER DIRECT COSTS: (itemize)		
Equipment (including software, hardware, purchases/rentals):	$4,000 (1 year USGS CHS fees)	
Publication Costs:		
Office supplies, Training, Other Expenses (specify):		
Total Other Direct Costs:	$7,000	
Total Direct Costs:	$27,000	$40,000
Indirect Costs (%):		
GRAND TOTAL:	$27,000	$40,000