**Developing a workflow for managing large hydrologic spatial datasets to assist water resources management and research**

# Project Charter

# Vision

Large, spatially distributed datasets have increasingly become more abundant, but there is currently no workflow that efficiently manages, analyzes and visualizes these datasets, ultimately dampening their usability and assistance in water resource management/research. As an example, repeat observations of spatially-distributed snow observations can assist real-time streamflow forecasting with higher reliability than current methodologies that use sparse point observations. NASA JPL used airborne LiDAR technology to collect high-spatial resolution (3m) snow depth datasets during the snowmelt seasons of 2013, 2014 and 2015 (the lowest snowpack on record) over Tuolumne River watershed, source of San Francisco’s water supply, to help address this problem. The dataset is composed of a bare earth model and snow depth spatial grid that is collected 6-11 times per year. We plan to use this extensive dataset with hydrologic modeling and three-dimensional meteorology output from a weather model (WRF, 1TB NetCDF file for one year) to complete various spatio-temporal analyses. In addition, our group is expected to receive in the near future (1-3 months) similar high spatial resolution (1-3m) datasets for the Olympic Mountains covering larger geographic areas, making each file about 20GB (we already have a sample 9GB file). We are also planning the acquisition of high-resolution thermal infrared imagery (1-3m) during January 2016 (about 3-5 GB each file) to use in conjunction with modeling other gridded weather data and satellite imagery. We expect the total amount of data to reach up to 10-16TB. **Within the incubator, we envision creating a workflow based on the existing WRF meteorology files (1TB spatio-temporal) and LiDAR-derived snow depth spatial datasets (1 and 9 GB) to be applicable to other existing or incoming data (**it isCRITICAL to figure out the 1-9 GB workflow on spatial fields before we have 20 GB files**). The workflow will help integrate high-resolution spatio-temporal datasets with hydrologic modeling to improve water resources management.**

# Objectives

Specific objectives we expect to achieve within the incubator:

1. ***Data storage/management/automation***: We will evaluate options for data handling and storage solutions for our growing data needs. We will explore Microsoft Azure and Amazon AWS cloud services for both data storage and data processing. These solutions will be optimized such that all research group members can access and efficiently process the datasets in the cloud and on local machines. We want to understand the longer term cost structures/estimates for each option.
2. ***Data processing and manipulation*.** We will develop scripts to process and manipulate the data according to our research needs. Because our datasets are in various formats (binary, NetCDF, geotiffs etc) we will identify the most appropriate tools to convert and process the data in a unified workflow based mostly on Python and Matlab scripts and NetCDF data format. We will explore processing environments and software both on local machines and in the cloud.
3. ***Visualization***. We will explore Python tools for visualization of large datasets in netCDF format as well as web-based interactive visualization solutions***.***
4. ***Reproducibility and Open Science***. We will explore solutions for data storage for scientific and public access (configuration, local/cloud storage and visualization).

# Success criteria

* Identify the most appropriate cloud service for data storage as a function of data size, number of people using it, frequency of access and cost of using software licenses for data processing in the cloud
* Develop scripts and workflow for data processing to speed up scientific output
* Identify the most appropriate visualization tools for the large spatio-temporal datasets

# Deliverable Schedule

Jan 26th : Preliminary exploration of cloud solutions completed. Python xray visualization solutions tested.

Feb 15th : Testing of Matlab data processing in the cloud completed. Cloud computing evaluation continued.

Feb 29th: Cloud storage and data processing solutions finalized.

Mar 5th : Web-based visualization solutions and project wrap-up.