1.0 Project Summary

Name of Project: Scalable Serverless Workflows for Processing Cubesat Imagery to Identify Flowering Hotspots in Sub Alpine Meadows

Project Lead and Contact Details: Amanda Tan, eScience Institute, University of Washington (amandach@uw.edu)

Project Partners and Contact Details: Aji John, Dept. of Biology, University of Washington (ajijohn@uw.edu); Kristiina Ausmees, Dept. of IT, Scientific Computing, Uppsala University, Sweden (kristiina.ausmees@it.uu.se)
Proposed start and end dates: Aug. 2019 - Jan. 2020

Budget Requested: $8500
Budget Summary: Refer to Section 3.0 for Deliverables

- Develop workflow on AWS ($3500)
- Test and validate workflow with Planet imagery and in-situ data ($2500)
- Conference travel, presentations and reports to the 2020 ESIP Winter meeting ($2500)

2.0 Project Outline

Project Description: Dramatic shifts in phenology in response to climate change have been observed for numerous species (Parmesan 2006; CaraDonna et al. 2014). Montane ecosystems are particularly at risk as they are most susceptible to climate warming (Chen et al., 2011). Alpine wildflowers that are dominant in these ecosystems have been shown to be experiencing local level extinction because they are highly sensitive to spring and summer temperatures (Inouye, 2008; Theobald et al., 2017; Panetta et al., 2018). However, the effects of warming on these alpine wildflowers are difficult to quantify due to a paucity of long term studies (Theobald et al., 2017).

Traditionally, discernible phases of plant phenology have been collected by routine field measurements during the growing season. While collected data are highly valuable in documenting life stages and the environmental growing conditions of these species, the long-term viability of these experiments are less certain (Kobori, 2016). Satellite-based imagery (e.g. Landsat 8 and Sentinel) have been key for looking at spectral signatures to detect changes in plant phenology (e.g. Bradley, 2013; Kobayashi et al., 2018) but its main limitation had been a coarse spatial resolution (10m - 30m) with low temporal frequency (biweekly to once a month). However, the advent of CubeSat imagery that provide daily 3 - 5m resolution imagery (e.g. Planet Labs, Digital Globe, etc.) have opened up the possibility of detecting the phenological changes with much more accuracy and sustainability (Cooley et. al. 2019).

The goal of this project is to develop an *workflows using serverless platforms* to process satellite imagery at scale and to derive models for detecting peak flowering phenology from high-resolution CubeSat imagery. In this project, we will use a study system at Mt. Rainier National Park (MORA) that was established by Hille Ris Lambers Lab at the University of Washington and contains an extensive set of past phenological and climate records for alpine wildflowers (Ford et al., 2013; Theobald et al., 2016; Theobald et al., 2017). Raw spectral information along with derived indices will be used to detect the extent and phenology of alpine flowers. We will then validate the predictions with the aforementioned field-collected flowering phenological data and against flowering observations gathered by a citizen science program called MeadoWatch (http://www.meadowatch.org). The goals will be achieved through the objectives outlined in Section 3.0 below.

3.0 Project Objectives, Significance and Impact:

Objective 1: Develop an experimental cloud-based, serverless workflow for retrieving and analyzing satellite data

We propose to develop an experimental workflow that utilizes *serverless computing* to process and analyze satellite data, specifically from Planet Labs. Current scientific workflows encompass a wide array of disparate tasks that vary in execution time and requires extensive computational infrastructure setup and maintenance (e.g. using traditional storage options and setting up virtual machines or computers with enough processing speed and power). Serverless computing or Function as a Service (FaaS) can help reduce cost overruns from inefficient resource use by only running code and eliminating the need for managing infrastructure.The modules of a satellite serverless workflow would consist of user-defined functions i.e. executable code written in any supported language (e.g. Python, Java, etc.) that are deployed to a cloud provider that handles the allocation of resources and execution. Individual functions can be combined into complex workflows by defining rules that specify the order in which they are to be executed, describing inter-task dependencies, and outlining how outputs and inputs are passed between tasks.

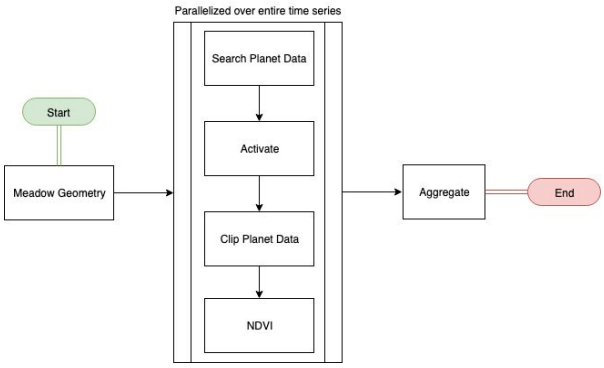
Objective 2: Can Planet satellite imagery be used to develop models for detecting peak flowering phenology?

Use of spectroscopy in remote sensing has been proven to be very informative in assessing changes in an area over time (Baumann et al. 2017). Green chromatic coordinate (Gcc), and normalized difference vegetation index (NDVI) are estimates of greenness, and has been established as a reliable measure for tracking vegetation color (Schwartz. 2013). Remote sensed imagery when fused with environmental and physiographic information might help in identifying flowering signal. We will combine past manual observations of peak flowering in meadows (i.e. dataset by Theobald et al., 2017) with Planet imagery, climate data from MORA weather stations and physiographic data extracted from 1-m digital elevation model provided by

the Puget Sound Lidar Consortium (http://lidarportal.dnr.wa.gov/) to develop, train and validated peak flowering models.

4.0 Key Project Steps and Timeline

a) Develop serverless workflow on AWS (Aug. 2019 - Oct. 2019)



We will implement a satellite imagery analysis pipeline in terms of a workflow executable on Amazon Web Services (AWS) Lambda, with the possibility of generalization to other providers such as Google Cloud Platform Cloud Functions and Microsoft Azure Functions. However, for this proposal, we will focus our architecture only on AWS. Defining our workflow in the FaaS context will include writing all functionality in terms of tasks in Python (<300Mb in memory) that can be executed in less than 5 minutes, and defining rules for their orchestration. Figure 1 shows the tasks that will be performed in retrieving, clipping and extracting NDVI information from Planet data. Each box represents a task (or Python code) that will be executed using AWS Lambda. The tasks will be defined as to ensure fully distributed operation, with intermediary data storage, logging and inter-task I/O handled by use of cloud-based object storage, e.g. AWS Simple Storage Solution (S3).

Figure 1. Example workflow highlighting the tasks involved in using Planet imagery for running NDVI across agiven set of meadows. The double wall boxes signify parallelization i.e. running the task over all the available dates.

b) Develop flowering phenology models using Planet satellite imagery (Oct. 2019 - Dec. 2019)

The NDVI and Gcc index obtained from Project Step 1(a) will be used along with climate and physiographic data to develop a peak flowering models.. In developing the peak flowering model, we will explore two regression-based algorithms (i.e Logistic and Principal Component) and two machine-learning based models (Random Forest and Gradient Boost) to as they have shown evidence of capturing flowering phenophase (Czernecki et. al. 2018). Existing observations which are for 6 years would be split to train and validate the models. The accuracy of the models will be evaluated using root-mean-squared-error (RMSE) and F-measures.

c) Bundling algorithms for publication, preparation of materials for ESIP Winter 2020 meeting (Dec. 2019 - Jan. 2019)

Significance and Impact: The serverless workflows that we propose accelerates computationally-intensive research by reducing the need to deploy and maintain cyberinfrastructure. The ability to parallelize elements of the workflow over multiple Lambda (serverless) requests also facilitates a more efficient use of time. While this proposal focuses on a workflow for processing satellite imagery, this workflow can be co-opted into other domains, from forest fire ecology to genomics. The development of algorithms and machine-learning models in flowering phenology detection would be key in furthering research in change-detection at a finer scale. Additionally, establishing a reproducible pipeline to fetch satellite imagery and evaluate it in tandem with other imagery providers may help the adoption and use of satellite imagery in climate change driven initiatives.

5.0 Outreach

The proposed serverless workflow is of benefit to a larger community beyond Earth Science. We will actively explore possibilities with our project partners to expand our workflow to other domains; this can be achieved by holding a training session at the ESIP Winter 2020 meeting, producing detailed and robust documentation as well as ensuring all algorithms are fully reproducible through containers and AWS machine images (AMI) and allowing open access to all data produced. We will also promote our work further by submitting a paper to a peer-reviewed journal (e.g. Environmental Research Letters) and submitting an abstract to the American Geophysical Union Fall Meeting 2019.

6.0 Integration with the ESIP community

We will work closely with the team from the ESIP Winter 2019 Incubator project ("Developing Workflows for Assessing High-Resolution CubeSat Imagery to Infer Detailed Snow-Covered Areas for Studying Changes in Ecosystems and Water Supply") to perform synergistic activities such as model development and scientific analyses. Further, we will convene a workshop at the ESIP Winter 2020 meeting to solicit feedback on the tools develop and to teach Earth Science participants about scientific workflows in serverless architecture.

7.0 Project Partners (if applicable)

Dr. Amanda Tan will work on the outreach, scientific questions and sustainability design of the program while graduate student Aji John and project partner Kriistina Ausmees will drive the development and implementation of the serverless workflows. We will also collaborate closely with Dr. Nicoleta Cristea (UW Civil and Environmental Engineering) and PhD candidate Tony Cannistra (UW Biology) on expanding the potential of the tools for scientific analysis and Dr. Ka-Yee Yeung (UW Tacoma Computational Biology) will assist in expanding the architecture beyond the Earth Science domain.