Scientific Data Analysis Web Application

CSCI 4308 Senior Capstone

Seongmin Choi, Robert Crimi, Connor Guerrieri, Bo Han, Hannah Keller, and Hannah Thomas

Project Overview

Over the past decade it has become increasingly clear that the Earth's climate is changing rapidly. Governments and private institutions have progressively invested more resources into the study of the impacts of climate change, but the are many barriers that currently slow down the advancement of scientific understanding. Due to the variety of computer simulation models, data formats, and data analysis languages, one of the largest barriers facing the progression of scientific understanding is reproducibility, or the ability to reproduce and verify the results of another's analysis. The goal of our project is to create a python web based application that not only allows users to easily build and run data analysis workflows, but do so without the need of specific programming and climate analysis knowledge.

Current Problem

The goal of the climate modeling community is to provide reliable, accurate, and credible information to those who make decisions on how or whether to respond to climate change. Two of the largest barriers currently slowing down the advancement of scientific understanding are the expertise needed to run analyses and the inability to effectively share and reproduce these analyses. Climate scientists interested in expanding their knowledge not only need to be experts in their scientific field, but also in the installation and configuration of software, translating data formats, and variety of analysis tools. Impact users, decision makers outside the climatology field that are invested in the impacts of climate change, such as city planners, utility managers, or park rangers, need easy access to climate data and a way to analyze this data without scientific expertise. Not only are these users facing challenges to analyze and access climate data, but there is specific expertise around handling the data, such as understanding model interactions and projections. For both scientists and impact users, being able to access and analyze climate data as it stands today requires too much overhead to make it an effective process.

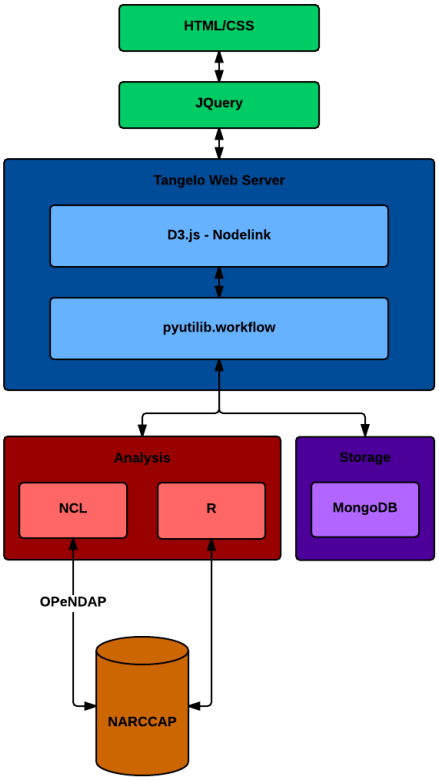
The Solution

The overall goal of the project was to break down the different aspects of creating a climate data analysis workflow into steps that were easy-to-use, reusable, and easily shareable. Our solution is a python based web application that allows users to dynamically create data workflows by connecting different pre-packaged analysis steps. To make it intuitive for a user to keep track of their workflow, we incorporated a visualization that displays each step and their connections to other steps. As each step is added, the workflow as a whole is run and

the user is given the option to download the data created after the latest step. Lastly, users are able to save their workflows and are given a serial number for that workflow. Using the serial number a user can either upload that workflow and update or share the serial number with another user to access.

The application is using a python based web framework, Tangelo, which was built specifically to support agile data management and visualization. To structure the workflows on the back end, we modified a python workflow library called pyutilib.workflow, which creates workflow objects containing multiple tasks, or steps. Each step available for the workflows are either NCL, NCAR Command Language, or R scripts which perform the actual analysis on the data and are called through python. For the scope of the project we focused on using NARCCAP, North American Regional Climate Change Assessment Program, data in NetCDF format, a common climate data format. After each analysis step, a new NetCDF file is created and the filename is then passed between steps and becomes available for the user to download. As steps are added to the workflow, the workflow library sends an updated workflow to a plugin in Tangelo, called Nodelink, that dynamically updates the visualization seen by the user. Workflows and their corresponding output files get saved in a MongoDB database via another Tangelo plugin, making them available for uploading or sharing given a corresponding serial number.

Through python we were able to run a web server, utilize different analysis languages, and support visualization to create an application for easy build out of scientific data analysis workflows that can be reused and shared.



Future Steps

This project's purpose was a proof of concept for a team at NCAR with the idea of expanding on the platform we created. The idea is to add access to more databases other than NARCCAP, add more analysis steps and steps that utilize more analysis languages, and create a template in which users in the future are able to create the own steps and submit them to be added to the application.

Contact Information

For more information on the project, please contact:
Brian Bonnlander, NCAR: bonnland@ucar.edu
Seongmin Choi, Deployment Lead: Seongmin.Choi@colorado.edu
Robert Crimi, Architecture/Research Lead: Robert.Crimi@colorado.edu
Connor Guerrieri, Source Control Lead: Connor.Guerrieri@colorado.edu

Bo Han, Test Lead: boha4482@colorado.edu
Hannah Keller, Team Lead: Hannah.Keller@colorado.edu
Hannah Thomas, Documentation Lead: imaginationandtech@gmail.com