# Using Geoweaver to Make Snow Mapping Workflow FAIR

Ahmed Alnaim (Alnuaim)
Center for Spatial Information Science and Systems
George Mason University
Fairfax, USA
aalnaim@gmu.edu

Ziheng Sun
Center for Spatial Information Science and Systems
George Mason University
Fairfax, USA
zsun@gmu.edu

***Abstract*—This paper replicates a snow depth estimation workflow in Geoweaver based on data from NASA National Snow and Ice Data Center (NSIDC), Airborne Snow Observatory (ASO), and Moderate Resolution Imaging Spectroradiometer (MODIS) datasets to make it more FAIRable for the community. The workflow searches and retrieves spatiotemporally coincident data from various data sources, preprocesses the data, and trains/tests it using a linear regression model and a deep learning model. The trained model will be able to estimate snow depth based on new remote sensed observations. Our experiments show that Geoweaver can significantly improve the capacity of sharing and synchronizing workflow among team members or individuals from other teams. Using the simple one-stop import/export button, scientists can share not only source code, but also all the history of activity and issues during their work so the people on the receiving end can see those struggles and avoid them.**

***Keywords—cyberinfrastructure, provenance, geospatial artificial intelligence, remote sensing, machine learning workflows***

## I. Introduction

Snow scientists have been struggling with sharing their work or reusing experiments from other colleagues. The lack of raw data, problems with authentication, storage capacity, computing resources, environment discrepancies and incompatibilities, as well as misconceptions about the same processing procedures among individuals, are all major contributors to the seriousness of the reusability and replication issues [1] in the snow community. To address the FAIR challenge, we experimented with Geoweaver as a promising solution. Geoweaver is an open-source workflow management software that lets users set up scripts and link them into traceable and manageable workflows [2]. The workflow incorporates user-created processes that lets the user operate it and any underlying processes locally or remotely on different hosts. Geoweaver simplifies the reusability aspect of building a software-based research experiment by allowing the creator of the workflow to export all of the necessary parts so that another user may quickly load it into their Geoweaver instance and replicate the results or edit the code. Geoweaver supports Python code, as well as shell scripts and Jupyter notebooks. Any of these languages/formats can be used to create a process that can be run independently or as part of a workflow. Snow depth has been researched for many years and is one of the primary markers that cryologists watch [3]. The snow workflow example used in this paper will automate a snow depth estimation workflow to train two models: a linear regression and a deep learning convolutional model to predict snow depth within a region in Colorado. This workflow contains code for collecting and searching the NSIDC and NASA EarthData, processing the data, feature engineering, and training and validating the model's accuracy.

## II. Background

More than one-sixth of the world's population (~1.2 billion people) relies on seasonal snowpack and glaciers for their water supply [4]. To understand the time and space variation in the snow's energy and mass balances along with the extensive feedbacks with the Earth's climate, water cycle, and carbon cycle, it is critical to accurately measure snowpack. SnowEx was initiated in the 2016-2017 winter with a field campaign in Colorado that was designed to evaluate the sensitivity of different snow remote sensing techniques [4].

## III. SnowEx Workflow

NASA Terrestrial Hydrology Program started the SnowEx initiative (THP), with a focus on articulating satellite remote sensing tactics and requirements. To access the NSIDC data that corresponds to the area defined in the polygon, a polygon area must be provided in the workflow through the SnowEx17 Ground Penetrating Radar, Version 2 dataset [5]. The polygon utilized in the workflow is a representation of the Grand Mesa study site in Colorado. To find information that matches the location and time in other data sources, the workflow will programmatically query additional data sources and extract information that matches the place and time supplied. The ASO L4 Lidar Snow Depth 3m UTM Grid, Version 1 is one data source that is gathered. The MODIS/Terra Snow Cover Daily L3 Global 500m SIN Grid, Version 6 is another automated data source gathered in the workflow. The workflow will start by reading a day's worth of SnowEx data and will collect ASO and MODIS coincident data. The data points are preprocessed and subsetted to match the SnowEx data and will be passed to Geopandas to provide point geometry. A concise data flow of the experiment is shown in Figure 1.

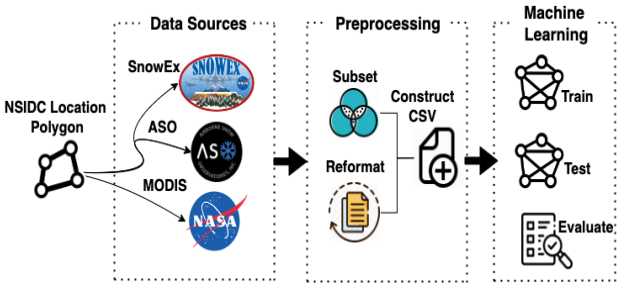


Fig. 1. SnowEx experiment dataflow

The workflow is split between 5 python files that contain all the code to automate the SnowEx experiment. To ensure reproducibility, replicability, reusability and provenance, across environments and potential future users, we can add each of those files in Geoweaver as a "Process". Creating a process in Geoweaver was simple and needed a name for the process, the language of choice, and the code itself. To manage all of those processes and have them monitored and automated by Geoweaver, we can create a workflow in Geoweaver that allows us to connect individual Python files (processes) to execute the overall workflow either locally or remotely on host(s). We compose the processes into a pipeline using the "Weaver" component in Geoweaver. Since the SnowEx workflow must be conducted in a specific order, the workflow can be constructed in a sequential manner to guarantee that all ensuing nodes (processes) are correctly carried out. Figure 2 shows the entire workflow for reference.

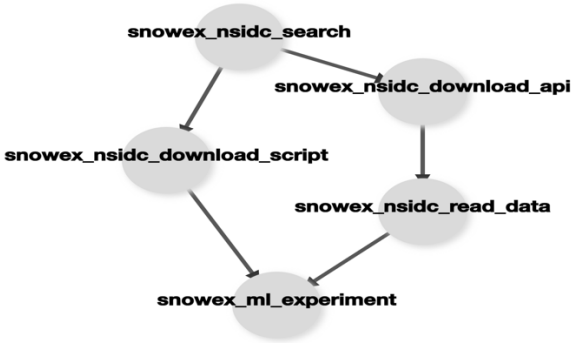


Fig. 2. SnowEx workflow on Geoweaver

## IV. Results

The workflow successfully produced evaluation metrics for the performance of the Linear Regression and Deep Learning models, replicating the NSIDC experiment on Snow Depth and Snow Cover. The final results for both models were an RMSE of 0.47 for the Linear model and 0.31 for the Deep model. The R2 score for the Deep Learning model was 76.22%, however it was just 47.74% for the Linear Regression model. Regarding the reusability, Geoweaver enables the team to collaborate seamlessly on a single project, prevents miscommunications and file losses, and automates the tracing and merging history for all processes on separate user's machines. Eventually, all the code and history are stored in Github via the export button in Geoweaver. The GitHub repo of the SnowEx Geoweaver workflow is publicly available [6].

## V. Discussion

Based on our experience with Geoweaver, we are confident in promoting Geoweaver as a useful productivity tool for research groups who do a lot of ad hoc scripting and repeated experiments with small changes in each iteration. It will significantly improve the reusability of research workflows and contribute to more consistent experiments. In our SnowEx use case, the workflow history can be one of the most significant assets for future researchers, even more valuable than the most recent code because it allows new researchers to pick up where we left off via saved outputs. Different research experiments have already been conducted by utilizing some or most of its processing through Geoweaver [7]. Geoweaver was effectively integrated as a platform for ad hoc iterative testing in a research that examined machine learning and remote sensing to quantify the accuracy of predicting NO2 for power plants in the United States [8]. Other research endeavors are currently being entirely developed on Geoweaver, such as crop mapping [9] and CMAQ-AI. This workflow is being created to have an alternative method for forecasting O3 (Ozone) simulation data from the Community Multiscale Air Quality Modeling System (CMAQ).

## Acknowledgment

This project is funded by NASA ACCESS #80NSSC21M0027, NSF Geoinformatics program (EAR-1947893 & EAR-1947875, 2020), NSF Cybertraining (#2117834, 2021).

## References

[1] Z. Sun, L. Sandoval, R. Crystal-Ornelas, S.M. Mousavi, J. Wang, C. Lin, N. Cristea, D. Tong, W.H. Carande, X. Ma, Y. Rao, etc, "A review of earth artificial intelligence," Computers & Geosciences, p.105034, 2022.

[2] Z. Sun, L. Di, A. Burgess, J. A. Tullis, and A. B. Magill, "Geoweaver: Advanced Cyberinfrastructure for Managing Hybrid Geoscientific AI Workflows," ISPRS International Journal of Geo-Information, vol. 9, no. 2, p. 119, Feb. 2020, doi: 10.3390/ijgi9020119.

[3] S.G. Warren, I.G. Rigor, N. Untersteiner, V.F. Radionov, N.N. Bryazgin, Y.I. Aleksandrov, and R. Colony, , "Snow depth on Arctic sea ice," Journal of Climate, 12(6), pp.1814-1829, 1999.

[4] NASA SnowEx, "About the Mission | SNOW." https://snow.nasa.gov/campaigns/snowex/about (accessed Jul. 16, 2022).

[5] R. Webb, D. McGrath, K. Hale, and N. P. Molotch, "SnowEx17 Ground Penetrating Radar, Version 2." NASA National Snow and Ice Data Center DAAC, 2019. doi: 10.5067/G21LGCNLFSC5.

[6] Geoweaver team, "snowEx-geoweaver", 2022. Accessed: Jul. 18, 2022. [Online]. Available: https://github.com/earth-artificial-intelligence/snowEx-geoweaver

[7] Z. Sun, L. Di, and H. Fang, "Using long short-term memory recurrent neural network in land cover classification on Landsat and Cropland data layer time series," International Journal of Remote Sensing, vol. 40, no. 2, pp. 593–614, Jan. 2019, doi: 10.1080/01431161.2018.1516313.

[8] A. Alnaim, Z. Sun, and D. Tong, "Evaluating Machine Learning and Remote Sensing in Monitoring NO2 Emission of Power Plants," Remote Sensing, vol. 14, no. 3, p. 729, Feb. 2022, doi: 10.3390/rs14030729.

[9] Z. Sun, L. Di, H. Fang, and A. Burgess, 2020. Deep learning classification for crop types in north dakota. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, pp.2200-2213.