A Geospatial-Natural Language Processing Analysis of the Snow Albedo Literature

Project Details

Project lead: Eric A. Sproles, Montana State University (eric.sproles@montana.edu)
Project partners: Chaowei "Phil" Yang (cyang3@gmu.edu), Charles Gatebe (charles.k.gatebe@nasa.gov), Anne Nolin (anolin@unr.edu), Christopher Crawford (cjcrawford@usgs.gov), Paul Houser (phouser@gmu.edu), Kelly Gleason (k.gleason@pdx.edu)
Proposed start and end date: 1 Sept 2020 - 31 March 2021
Budget Requested: $9000
Deliverable 1 - Complete Natural Language Processing (NLP) of the snow albedo literature ($3000).
Deliverable 2 - Map the geographic results of the NLP analysis using Uber's H3 framework ($3000).
Deliverable 3 - A collaboratively written assessment of results from the NLP and geographic analysis that identifies regions and topics well and less represented in the published literature ($1500).
Deliverable 4 - Plan travel for ESIP Summer Meeting 2021 ($1500)

Project Outline

Project description and objectives*:* Snow albedo has long been recognized as a key variable for snow hydrology, climate modeling, and energy balance calculations. Collective research efforts have generated an increasing number of scientific publications on snow albedo, however a comprehensive examination of the published literature on this topic does not yet exist. Since conducting literature reviews is so time-consuming and tedious, this project will integrate new methods by: 1) completing a Natural Language Processing (NLP) analysis of the published literature on snow albedo; 2) transitioning the NLP results into a scalable mapping analysis; and 3) producing a summary report of the topical and geographic breadth of the snow albedo literature.

These efforts directly support the goals of Snow Albedo Working Group (SAWG) to complete a scoping document and accompanying journal article focused on previous studies and future directions of snow albedo research. The SAWG is composed of 20 domain science professionals from across government agencies and academia in the United States and Europe.

The primary objective of the SAWG scoping document is to provide a comprehensive review of published literature on the topic of snow albedo indexed in digital libraries (e.g. the IEEE Xplore, Web of Science, JSTOR, PubMed, and Springer), synthesize and map the results, and identify key knowledge gaps. The methods to evaluate a discipline-focused analysis requires researchers to organize, read, and provide meta analysis on the body of literature - a process that requires a considerable investment of human capital [1]. Machine Learning, specifically Natural Language Processing, expedites the process, organizes the results, and provides a replicable framework for similar or continuing efforts. A NLP

analysis also has the ability to extract geographic characteristics, which for Earth Science research is particularly applicable and identify regions where the state of knowledge in snow albedo science is missing (e.g. Andes, Atlas).

NLP Methods: Automatic text understanding is a NLP task of programatically interpreting the meaning of text by a machine. In this project, keywords provided by the SAWG will serve as the analysis targets in the abstract and conclusion sections of the body of literature. Open-source Apache Tika code readily detects abstract and conclusion sections from a research article, and automatically extracts contents from multiple file formats, e.g., pdf, word. Open source NLP tools (e.g., Stanford NLP packages) recognizes name entities such as organizations, person names, locations, percentages, time expressions, and quantities from extracted contents. Additionally, domain specific terms can serve as training data for the NLP to extract entities and their relationships, key findings, and future research from abstract and conclusion.

The geographic information of the study area will be extracted using named entity recognition packages and Geocoding APIs will be leveraged. For a given paragraph, open-source and well-built Named Entity Recognition (NER) tools can automatically recognize location from text. Meanwhile, geographically related keywords can serve as input to train a customized named entity recognition model to detect domain-specific geographic information. In addition, Geocoding APIs automatically detect geographic information from input text, providing geospatial supplements to the output of the named entity recognition methods. These efforts will be led by Dr. Chaowei Yang and Paul Houser at George Mason University.

Mapping Methods: Geospatial indexing provides a means to organize, aggregate, and analyze the NLP results at multiple scales. We will use the H3 algorithm to partition the Earth into hexagons at 16 different levels of detail. Originally developed by Uber, the open-source H3 framework transforms latitude and longitude into a unique index that can be redrawn at 16 spatial resolutions. Additionally H3 is hierarchical, facilitating single maps to be drawn at multiple resolutions. For example, in the case of snow albedo, the resolution of California's Central Valley would be rendered at coarse resolutions (large hexagons), while the Sierra Nevada would be at a higher level of spatial detail (small hexagons).

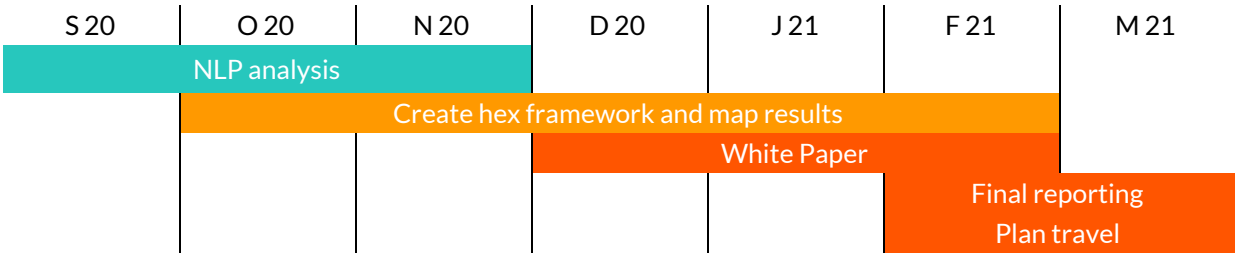
The geographic results of the NLP analysis, including key terms, will be transitioned into the H3 framework, mapped, and geospatially analyzed. For example, articles that contain the term "light absorbing particles" will have their geographic information extracted and mapped. Subsequent spatio-temporal analysis such as a Getis-Ord Gi* cluster analysis will identify high or low clusters of

research activity [2]. The resulting map will identify where research focused on "light absorbing particles" has occurred, and more importantly where geographically research gaps in snowy areas may occur. The NLP results and resultant cluster analysis will be visualized using MapBox, and hosted and maintained by Dr. Eric Sproles at Montana State University.

Assessment of results: The investigators on this project will collaboratively author a report on the topical and geographic breadth of snow albedo research based upon these results. The goal of this document is to identify underlying deficiencies in current research, better inform research design for subsequent efforts, and serve as a guiding document for program managers at different funding agencies.

***Project significance and impact**:* The SAWG is composed of early-to-senior career experts in measurement and modeling who are poised to make major scientific gains in responding to the full range of snow albedo science and applications. This group was organized at the request of NASA Terrestrial Hydrology Program to identify how and where to improve measurements and remote sensing of snow albedo, and how to transition these findings into a future proposal for a new satellite mission. In the 2018 *Decadal Survey for Earth Observation from Space* [1], snow albedo is considered a targeted Designated Observable, with a goal of characterizing climate forcings, more accurately determining snow melt rates, reducing uncertainty in snow/ice albedo feedback by a factor of two, and improving water balance calculations. This project will support a data science approach assessing domain specific literature, a process that is largely absent in the snow science and broader Earth Science community. This project will also support one graduate (George Mason) and undergraduate (Montana State) student, providing each the opportunity to work collaboratively and expand their professional network in the Earth Science community.

Description of key project steps and timeline:



Description of additional funding currently supporting this work: The majority of this work will be done in kind. Funding will primarily support student stipends and travel. SAWG does have a NASA Terrestrial Hydrology proposal submitted that will support a more detailed scoping document and a

workshop in 2021. If funded, the results of this study will directly complement and enhance the 2021 workshop.

Outreach

This project directly engages the SAWG - 20 Earth scientists that conduct research on snow albedo. In the near term, the findings will be implemented into SAWG planning and subsequent workshops to inform NASA Terrestrial Hydrology and federal agencies (USGS, NOAA) of its results. In the longer term, the project's impact will be judged on how it informs subsequent studies of snow albedo. This is not a stand alone project. The Geospatial-NLP analysis methods we develop will also be of value to other Earth Science disciplines and highlight how data science can augment traditional means of reviewing literature.

The source code, results, and knowledge generated by this project will be published on ESIP's GitHub account, in addition to the mapping efforts hosted on Sproles' website. This project will be presented to the SAWG and at subsequent professional meetings (e.g. ESIP, American Geophysical Union, Association of America Geographers, American Meteorological Society) by project investigators. This summary document will also serve as a forerunner for submission to a peer reviewed journal (e.g. Water Resources Research or The Cryosphere). The project team has representatives from NASA Goddard and the USGS EROS Center, facilitating for ready dissemination to colleagues at these centers.

Project Partners

Description of project partners (agencies/individuals) and their involvement: The project team is composed of investigators from two different federal agencies (NASA and the USGS) and four different universities. Additionally students from George Mason University and Montana State University will actively contribute to this project and be co-authors on any presentation or publications. The results will be presented to the SAWG and implemented into the group's scoping project and workshop.

How will this project engage members of the ESIP community: The integration of Data and Earth Science is the foundation of ESIP, and this project integrates open-source approaches to NLP and spatial analysis. This modernization of workflows will be deployed in a cloud environment that will be replicable through well-documented notebooks published on ESIP Labs' GitHub page. This software will provide a meta analysis of the scientific literature and incorporate the geographic information associated with the corpus of research. While the scope of this research focuses on snow albedo, a targeted Designated Observable, it will also be readily adaptable to assess other themes in the scientific literature.

References

1. Zdravevski E. et al. (2019) Automation in Systematic, Scoping and Rapid Reviews by an NLP Toolkit: A Case Study in Enhanced Living Environments. In: Ganchev I., Garcia N., Dobre C., Mavromoustakis C., Goleva R. (eds) Enhanced Living Environments. Lecture Notes in Computer Science, vol 11369. Springer.
2. Songchitruksa, P., & Zeng, X. (2010). Getis–Ord spatial statistics to identify hot spots by using incident management data. *Transportation research record*, *2165*(1), 42-51.
3. National Academies of Sciences Engineering and Medicine *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*; National Academies Press: Washington, DC, 2018.