Final Workshop Report

Building Upon the EarthCube Community- A geoscience and cyberinfrastructure workshop

June 27-28 2023
Information Sciences Institute, Marina Del Rey, CA.

Organizing Committee

Deborah Khider[1], Mike Daniels[2,3], Nick Jarboe[4]

Authors
Deborah Khider[1], , Mike Daniels[2,3], Nick Jarboe[4]

Authors - Breakout session recommendations
Bridget Hass [5], Arika Virapongse[6], Yuhan (Douglas) Rao [7], Daniel Fuka [8]

1. Information Sciences Institute, University of Southern California, 2. NCAR, 3. Ronin Institute, 4. Jarboe Enterprises, 5. National Ecological Observatory Network, 6. Middle Path Ecosolutions, 7. North Carolina State University, 8. Virginia Tech

Contents

1 Executive Summary

Over nearly a decade, NSF's EarthCube initiative has highlighted the importance of collaboration between geoscience and cyberinfrastructure experts, transforming geosciences research through improved data access and analysis. EarthCube funded 85 projects, leaving a legacy of valuable products and publications, and fostering a vibrant community participating in various activities. Recent sustainability reports have recommended actions to maintain EarthCube's momentum, including continued annual meetings. As a result, the "Building Upon the EarthCube Community: A Geoscience and Cyberinfrastructure Workshop" convened 46 geoscientists and cyberinfrastructure experts to catalyze new opportunities at the intersection of geosciences and cyberinfrastructure, promote cutting-edge science, showcase progress, and foster collaboration.

The workshop attracted 36 abstract submissions and featured morning plenary sessions with keynote speakers and 15-minute oral presentations and afternoon sessions shared between breakout discussions and poster presentations. The final day included a breakout discussion about the future of geoinformatics.

2 Overview of the workshop

Since the inception of NSF's EarthCube initiative, the geoscience community has learned the value of teaming up geoscience professionals with cyberinfrastructure experts. For nearly a decade, the EarthCube community has been transforming the conduct of geosciences research by developing and maintaining a well-connected and facile environment that improves access, sharing, visualization, and analysis of data and related resources to foster a better understanding of our complex and changing planet. The EarthCube program had an impressive outcome of 85 funded projects (EarthCube, 2022) over the course of its lifetime and leaves behind a legacy of excellent products and numerous publications (Maull & Mayernik, 2022). Most importantly, EarthCube's success has been in the development of a vibrant, engaged community as evidenced in their continued participation in activities such as the Federation of Earth Science Information Partners (ESIP) Council of Data Facilities (CDF) cluster, the American Geophysical Union (AGU), - including assistance in the building of a submission mechanism for executable science Notebooks-, and the EarthCube Council of Funded Projects meetings. After a long period of community information gathering capturing the most important elements of EarthCube sustainability, the continued enthusiasm and a path forward has been described in three recent reports on sustainability: 1) the EarthCube Sustainability Panel Report (Daniels et al., 2022), 2) the CDF Sustainability report (Jarboe et al., 2022), and 3) the Council of Funded Projects Sustainability report (Virapongse et al., 2022a). These reports call out a series of recommended actions that the community can take to continue the momentum of the EarthCube initiative. Specifically, recommendation 11 of the EarthCube Sustainability Panel

Report (Daniels et al., 2022) calls out the need for continued annual meetings to support the existing funded projects and the community. These recommendations led to the "Building Upon the EarthCube Community: A Geoscience and Cyberinfrastructure Workshop" to continue upon the momentum created by EarthCube.

The workshop, held at the University of Southern California Information Sciences Institute on June 27-28th 2023, brought together 46 geoscientists and cyberinfrastructure experts. The main goal of the meeting was to catalyze the geoinformatics community towards new opportunities beyond EarthCube at the frontier between geosciences and cyberinfrastructure. The workshop was organized to promote cutting-edge science endeavors, to showcase progress on current projects, to spark new collaborations, and to build capacity among its constituents.

2.1 Workshop Chairs and Organizers

The organizing committee consisted of community members who (1) have been involved with EarthCube as principal investigators for specific projects and with its governance, (2) have extensive experience in workshop development, and (3) have been involved in crafting sustainability plans for the initiative. The organizing committee established the agenda of the meeting.

The following individuals were members of the organizing committee:

- Deborah Khider - Chair, University of Southern California/Information Sciences Institute
- Mike Daniels - Co-Chair, NCAR and Ronin Institute
- Nick Jarboe - Co-Chair, Jarboe Enterprises
- Lynne Schreiber - Co-Chair, San Diego Supercomputer Center

2.2 Organization of the workshop

The workshop's format encouraged synergies, discoveries, and collaborations among experts in science and technology to advance research at the intersection of geoscience and cyberinfrastructure. The organizing committee solicited abstracts through geoinformatics community listservs for oral presentations, posters, demos (including Jupyter Notebooks), and working sessions. In total, we received 36 diverse abstract submissions from various geoscience areas.

Morning plenary sessions consisted of keynote speakers and oral presentations selected from the abstracts. The first keynote speaker, Dr. Chris Mattmann (JPL), highlighted work done at NASA for the science missions. The second keynote speaker, Dr. Weiwei Duan (USC ISI) presented her work on object detection in maps using deep learning methods. This work won the DARPA map feature extraction challenge earlier in the year. The oral sessions

featured 15-minute presentations on diverse geoinformatics projects, emphasizing early-career researchers and researchers from underrepresented groups.

On the last day, participants formed breakout groups to discuss the future of geoinformatics, focusing on lessons from prior efforts, infrastructure investments, sustainability, training opportunities, and diversity initiatives. The recommendations from this breakout session are outlined in Section 5.

Afternoon sessions consisted of concurrent working sessions, which were solicited from the community. Four working sessions were accepted into the program: (1) Expanding open-source tools for working with aerial remote sensing data in Google Earth Engine, (2) Working towards AI-ready Geoscience Data Repositories, (3) Project-based sustainability for Earth Science data infrastructure, and (4) Building upon the Internet of Things (IoT) projects. The specific recommendations from this working sessions are detailed in Section 6.

The final afternoon sessions were devoted to posters and demos, promoting discussions on ongoing projects among participants. An overview of the workshop schedule is presented in Section 4.

2.3 Acknowledgement

The organizers and the participants of the "Building Upon the EarthCube Community" workshop are very grateful to the National Science Foundation for funding under RISE 2315484, and especially Raleigh Martin, Program Director for Geoinformatics, who participated in the workshop.

3 Participants

A call for abstract for oral presentations, poster, demos, and working sessions was advertised through various community lists targeted towards this particular community. The deadline for abstract submission was set to May 8th.

Abstracts were accepted into the program if they fit the general theme of the workshop, with priority given to early-career researchers, researchers from underrepresented group in the computer and geosciences, and projects funded under EarthCube. Early-career researchers were further supported by travel grants to attend the meeting.

3.1 Participant List

Participants are listed in alphabetical order.

Agarwal, Khushboo - The University of Texas at Austin/ Texas Water Development Board
Alkaee Taleghan, Samira - University of Colorado Denver
Brandenberg, Scott - UCLA
Chakraborty, Sudip - UMBC
Clyne, John - NCAR

Crosby, Christopher - OpenTopography / EarthScope
Curcic, Milan - University of Miami
Daniels, Mike - NCAR/Ronin Institute
Deauna, Josephine Dianne - University of Hawai'i at Manoa
Denis, Caroline - NA
Duan, Weiwei - University of Southern California
Elipot, Shane - University of Miami
Eroglu, Orhan - UCAR/NCAR
Ferrario, Iacopo - JRC
Fletcher, Lydia - University of Texas at Austin
Fuka, Daniel - Virginia Tech
Guo, Tianjing - Georgia State University
Hass, Bridget - National Ecological Observatory Network
Jarboe, Nicholas - Jarboe Enterprises
Khider, Deborah - USC Information Sciences Institute
Kirkpatrick, Christine - UC San Diego
Knoblock, Craig - USC Information Sciences Institute
Kumar, Deepak - State University of New York at Albany
Landers, Jordan - USC
Ma, Marshall - University of Idaho
Martin, Raleigh - NSF
Mattmann, Chris - NASA JPL
Mayernik, Matthew - National Center for Atmospheric Research (NCAR)
McHenry, Kenton - University of Illinois @ Urbana-Champaign
McKay, Nick - Northern Arizona University
Moon, Seulgi - UCLA
Nandigam, Viswanath - University of California San Diego
Nelson, Ellen - University of Wisconsin– Madison
O'Keefe, Patrick - New Jersey Institute of Technology
Pandey, Chetraj - Georgia State University, Atlanta, GA, USA
Porter, Brent - University of Texas Austin
Quinn, Daven - University of Wisconsin–Madison
Rao, Yuhan (Douglas) - North Carolina State University
Ringuette, Rebecca - NASA Goddard
Valentine, David - UCSD
Virapongse, Arika - Middle Path EcoSolutions
Wu, Qiusheng - University of Tennessee, Knoxville; Amazon Web Services
Zhang, Jiyin - University of Idaho
Zhu, Feng - National Center for Atmospheric Research

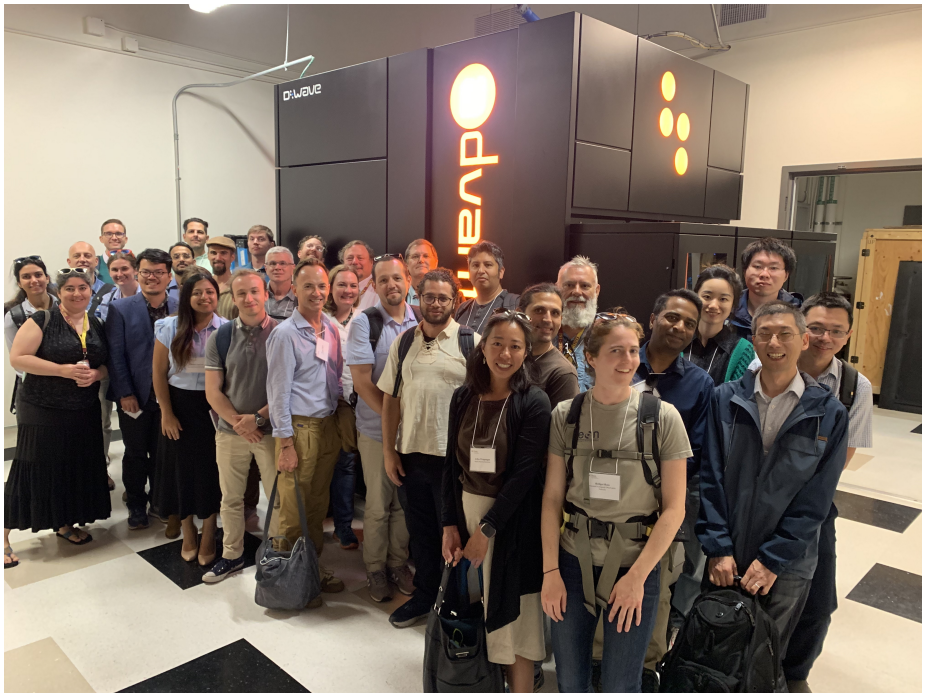

Figure 1: Photo of the workshop participants taken in from of the D-Wave quantum computer hosted at the Information Sciences Institute.

3.2 Demographics

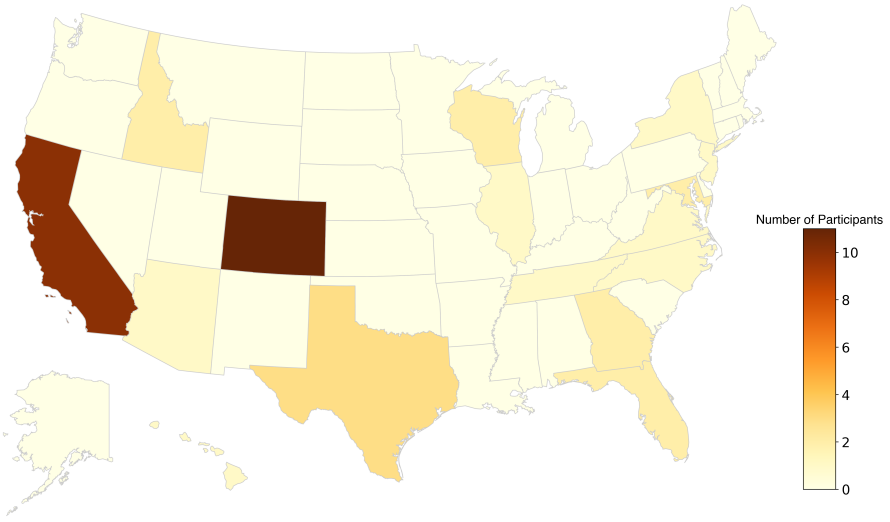

Figure 2: Number of participants per US States based on affiliation.

The workshop was primarily (54%) comprised of early-career researchers (graduate students, recent graduates and postdocs, and researchers within 10 years of their terminal degree, Fig 3).

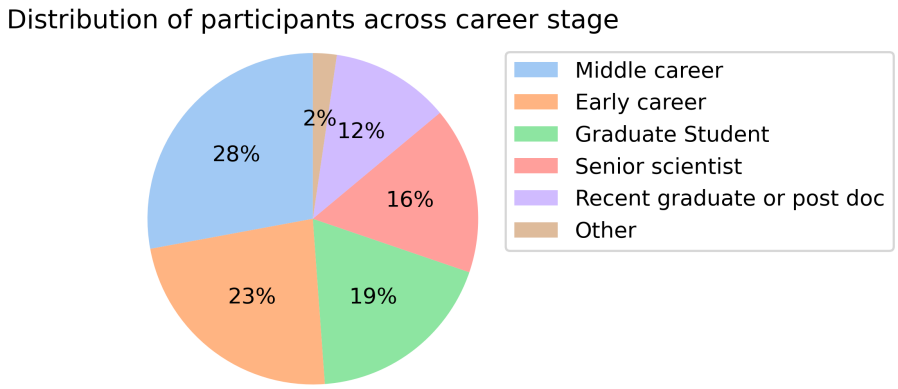

Figure 3: Proportion of participants that are current graduate students, recent graduates or post-docs, early career (within 10 years of terminal degree), mid-career (10-20 years after terminal degree), and senior researchers (more than 20 years after terminal degree).

4 Workshop schedule

The format of the workshop was designed to promote discussion among the participants and foster future collaborations. To this end, long breaks were scheduled throughout the day to encourage participants to discuss topics of interest.

On the first day, an early-career lunch group was organized to encourage these participants to get to know each other and start developing collaborations within geoinformatics outside of their or their PI's inner circle.

The days were scheduled to allow for plenary talks in the morning highlighting specific projects. Oral presentations were chosen to represent the diversity of geoscience applications, with priority given to early-career scientists. Two keynote speakers presented work done at NASA and DARPA. The keynote speakers were chosen to represent both the geoscience and computer science sides of geoinformatics, with one keynote reserved for an early-career researcher. Afternoons were reserved for working sessions that were proposed by participants followed by a poster and demo session to promote discussion about ongoing projects.

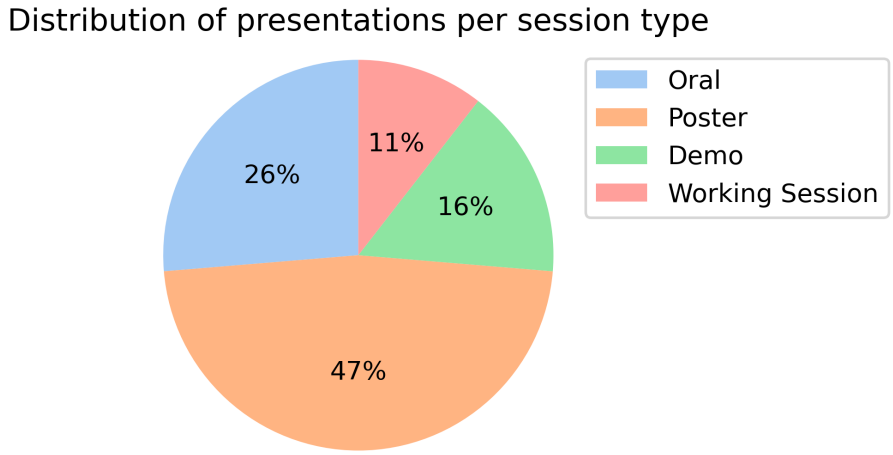

Figure 4: Proportion of session types in the workshop.

Tuesday June 27th
Morning Plenary - MDR1014

9:00-9:15 *Organizing Committee* Welcome, Code of conduct
9:15-9:25 *Craig Knoblock* Welcome to ISI
9:25-9:40 *Raleigh Martin* Update from NSF
9:40-10:20 *Chris Mattmann* Keynote I: AI and Machine Learning from Back of the Napkin Sketch to Rovers on Mars

BREAK

10:40-10:55 *Matt Mayernik* Recommendations for accessibility of simulation-based data
10:55-11:10 *Milan Curic* Challenges and Solutions Toward Efficient Lagrangian Data Analysis for Earth Sciences
11:10-11:25 *Ohran Eroglu* Project Raijin: Community Geoscience Analysis Tools for Unstructured Grids
11:25-11:40 *Deepak Kumar* Urban heat island variability assessment over transects for climate-sensitive sustainable urban heat island mitigation approach
11:40-11:55 *Scott Brandenberg* Assimilation of Earth Science and Geotechnical Engineering Data for GeoHazard Assessment

LUNCH

Working Session - MDR1014/MDR1016

14:00-15:20 *Bridget Hass* Expanding open-source tools for working with aerial remote sensing data in Google Earth Engine
14:00-15:20 *Yuhan (Douglas) Rao* Working towards AI-ready Geoscience Data Repositories

BREAK

Poster/Demo Session

15:40-17:00 *Poster*

Caroline Denis GeoSpecify collection management software
Arika Virapongse Sustainability of Earth Science data infrastructure projects
Ellen Nelson StraboSpot: A digital data system for collecting, storing, and sharing geologic data
Samira Alkaee Taleghan Sea Ice Classification by Learning from Label Proportions: How to Classify Sea Ice Type at Pixel-Level using Polygon-Level Ice-Chart Labels?
Patrick O'Keefe The Random Hivemind: An Ensemble Deep Learner. A Case Study of Application to Predicting All-Clear SEP Periods.
Sudip Chakraborty Black Carbon Atmospheric Rivers and Ice Sheet Melting: Machine Learning Approaches.
Tianjing Guo Towards a Computational Framework for Developing Data-driven Energy Policy
Chetraj Pandey Insights into Deep Learning-based Full-disk Solar Flare Prediction with Post hoc Explanation and Evaluation
Brent Porter Model Evaluation and Execution Platform: System architecture for model management and storage to support integrated planning and

interoperability at scale

15:40-17:00 *Demo*

Qiusheng Wu Interactive Geospatial Analysis and Data Visualization with Leafmap
Samuel Krasnoff VICTOR – A new Cyber-infrastructure for Volcanology
Shane Elipot A Lagrangian analysis workflow using the open source Python library clouddrift

Wednesday June 28th

Morning Plenary - MDR1014

9:00-9:40 *Weiwei Duan* Keynote II: Linear Object Detection in Map Images Using Transformer
9:40-10:20 *All* Brainstorming session on the future of cyberinfrastructure for the geosciences

BREAK

10:40-10:55 *Josephine Dianne Deauna* Improved grid-aware operations for ocean model analysis: expanding metrics functionality in the xgcm Python Package
10:55-11:10 *John Clyne* Update on Project Pythia: A Community Resource for Geoscientific Python Education
11:10-11:25 *Daven Quinn* Crossing rhe "collaborative barrier" for Macrostat and digital crust research
11:25-11:40 *Kenton McHenry* Ongoing Impacts from EarthCube Technical Efforts
11:40-11:55 *Seulgi Moon* Center for Land-Surface Hazards (CLaSH): Bridging Earth Data and Cascading Hazard Processes

LUNCH

Working Session - MDR1014/MDR1016

14:00-15:20 *Arika Virapongse* What's next?: Project-based sustainability for Earth Science data infrastructure.
14:00-15:20 *Daniel Fuka* Building Upon the IoT Projects within the EarthCube Community

BREAK

Poster/Demo Session

15:40-17:00 *Poster*

Venkadesh Samykannu Spatio-temporal trend analysis of satellite-based CHIRPS precipitation (1981-2017) and significance of crop production in selected crops over Tamil Nadu, India
Feng Zhu cfr: a Python package for climate field reconstruction
David Valentine The Quirks of Science on Schema: A Retrospective on Geocodes, and a Path Forward as Decoder
Julien Emile-Geay Pandas and the geosciences: a 4.5 billion year story
Alex Sun Building an AI/ML framework for flood inundation prediction
Lydia Fletcher Improving Traceability Throughout the Data Lifecycle: the DOLCE Approach to Provenance
Nicholas McKay A Paleoclimate Reconstruction Storehouse (PReSto) to integrate paleoclimate data, reconstruction workflows and visualization
Khushboo Agarwal Modeling electric grid vulnerability induced by natural events using Machine Learning and Geospatial analysis
Jordan Landers PaleoBooks: A Library of JupyterBooks for Paleoclimate Research

15:40-17:00 *Demo*

Jiyin Zhang OpenMindat Data API: the Automated Gateway to Two Decades of Crowd-Sourcing Mineral Data
Nicholas Jarboe Using the Modular FIESTA Software Stack for the Quick Stand-Up of FAIR Data Repositories for Drilling Core Multi-Sensor Track Data (CDR) and 40Ar/39Ar Age Data (KARAR)
Agbeli Ameko DEMO: Rural Community Engagement Through OpenIoTwx and EarthCube's CHORDS Platform

5 Recommendation for future directions

On the final day of the workshop, participants were asked to break into groups and discuss lesson learned from prior geoscience infrastructure endeavors, infrastructure investments, sustainability, training opportunities, and diversity initiatives. Each group wrote their recommendations on large notepads, which were collected at the end of the session for presentation and voting on the major ideas. The following represent a summary of the group discussion.

5.1 Lessons learned from previous geoinformatics endeavors

The importance of nurturing human infrastructure emerged as a common theme, emphasizing the need for annual meetings, outreach, educational initiatives, financial support, and highlight of early-career researchers. The group acknowledged the value of geoinformatics-focused workshops, whether held independently or as components within larger conferences like the AGU Fall Meeting. Additionally, participants noted the significant role played by initiatives like the NSF EarthCube program in promoting projects, facilitating connections between cyberinfrastructure experts and scientists, and managing the community effectively. One recommendation stemming from the discussions was to prioritize science-driven applications.

The participants also discussed the importance of a Central Office in the organization and implementation of the EarthCube initiative. Overall, the Office provided the "glue" that kept the initiative moving forward. From organizing annual meetings and other special purpose workshops, to hosting the website and document repository, to implementing governance initiatives, to sparking early career participation and more, the funding of an Office was absolutely critical to its success.

5.2 Infrastructure Investments

Recommendations for infrastructure investment fell into the following categories:

- **Support for intensive science:** As AI becomes more prevalent in the geosciences, computation will also become more intensive, requiring compute credits and cloud access, such as the one provided by the NSF ACCESS program.
- **Support for data services:**
  - The data needs to be live closer to where the computation will take place.
  - There needs to be an incentive for researchers to work with repositories to archive their data with the proper metadata while ensuring that credit for the original data producer is properly attributed.

  - Participants highlighted the difference between AI-ready data was (see Section 6.2) and analysis-ready data, which requires less metadata.
  - There also needs to be a focus on broad (i.e., heterogenous) data in addition to big data.

- **Training support in the use of HPC/cloud infrastructure**. Some researchers will need data help/support before they can use these systems. Furthermore, this training should occur in a core curriculum that focuses on Earth Science-related modules.

5.3 Sustainability

Participants recognized that the main barrier to sustainability is financial. Solutions offered included series of grants, advertisement-based income, donations, and subscription models. These funding sources could either be private or public. Participants also highlighted the need for more guidance from funding agencies in terms of what sustainability is and what it should look like. Finally, an emphasis on community governance for these projects was presented as a means to long-term sustainability of the tools/community.

5.4 Training Opportunities

Beyond training opportunities on HPC/cloud systems described above, training in software development and management is also valuable. Participants highlighted the fact that training on these broader skills, beyond the immediate field study, is valuable for any career. In addition, participants agreed that we need to support modern ways of training researchers, including, for instance, courses on online platforms such as Coursera.

5.5 Diversity Initiatives

Participants discussed the dimensions of diversity, which included background interest and underrepresented populations. There was also a discussion about not confusing diversity with inclusivity.

The discussion around diversity initiatives mainly focused on early-career scientists. The participants recommended more funding opportunities for early-career scientists to attend conferences, kickstarters/pilot project with a small amount of funding ($\leq$\$20k) as well as cohort-focused programs. Ideally, these programs will include PhD and MS students as principal investigators on projects and/or purses to senior PIs to organize grant competitions for students and programs within a program.

6 Specific breakout group recommendations

6.1 Session 1: Expanding open-source tools for working with aerial remote sensing data in Google Earth Engine.

This session, led by Bridget Hass from the National Ecological Observatory Network (NEON), provided an overview of NEON's Aerial Observation Platform (AOP) data and discussed the ongoing integration of AOP data into Google Earth Engine (GEE) as a publicly accessible dataset. Hass also demonstrated the use of GEE tools for AOP data, showcasing examples in both the Code Editor (JavaScript) and Python with `geemap`. The materials presented in this introduction is available here.

The resulting discussion centered around two main questions:

- How can NEON make the data more accessible, user-friendly, and inclusive?
- How can NEON/AOP data be better integrated with other complementary datasets?

While the workshop did not yield a general consensus, the discussions were valuable. One particular challenge, not unique to NEON but shared by other projects, is transitioning to cloud storage and cloud computing while ensuring that end-users will not be forced to use one cloud provider or another (for instance, the NEON team has moved storage to Google Cloud, but the tools remain cloud agnostic). Feedback was gathered during the working session, but this challenge remains unsolved.

During the working session, Quisheng Wu, a leader in GEE and Python, provided valuable feedback. Concerns were raised about adding open data to GEE, given that it is not entirely open-source due to Google's development and recent commercial use charges. However, GEE has gained popularity in the remote sensing community. It is reassuring to note that Google will continue to offer GEE for free in research applications, and they have taken ownership of the Python API and the `geemap` package, with plans for expansion.

6.2 Session 2: Working towards AI-ready Geoscience Repositories

Led by Yuhan Rao, the working session aimed to gather community input for enhancing data management practices in geoscience data repositories, ensuring they can meet the increasing demands of data-intensive research in the field. The session aimed to bring together data users and geoscience data facility staff to share perspectives from both sides around the topic. The focus of the session was on gathering participants' inputs and discussion via an online platform (Padlet). The discussion were carried out in breakout groups and focused on the following topics:

1. Understanding what "AI-ready" means for data users and data repositories.

2. Identifying the impact of the growth of AI and data science for geoscience data repositories.

3. Envisioning the actions that geoscience data repositories can do to support data-intensive sciences.

6.2.1 Understanding what "AI-ready" means

In the discussion, it became evident that participants held varying definitions and concepts of AI-ready data, often intersecting with the notion of "analysis-ready data," which also lacked a clear definition. Nevertheless, certain common themes began to emerge. Machine readability and usability were identified as crucial attributes influencing data readiness for data-intensive applications. This observation aligns with the ongoing adoption of FAIR data principles within earth and space sciences, which includes adherence to established community metadata standards.

Furthermore, participants emphasized the significance of data accessibility in facilitating AI and data-intensive research within geoscience data. In this context, accessibility encompasses various factors, including delivery methods, data formats, and supportive tools that aid users in accessing and utilizing the data. The choice of delivery methods for geoscience data can significantly impact the effectiveness of applications and research development, with cloud-based delivery methods gaining popularity due to the growing volume of geoscience data and community-driven initiatives like Pangeo Forge. Additionally, participants stressed the importance of utilizing standard or widely accepted data formats to enhance the usability of geoscience data, as these formats often benefit from robust support and community-developed tools developed for AI application development.

Lastly, participants also put forth specific areas where the concept of being "AI-ready" is driven by particular use cases or method-specific requirements. For instance, they discussed the need for labels or targets in supervised machine learning techniques and the management of time series data for recurrent neural network algorithms (e.g., LSTM). These cases often demand specialized data pre-processing for AI development and may have distinct considerations.

6.2.2 Identifying the impact of the growth of AI and data science

The groups then discussed the impact of the growing number of applications of AI and data science for geoscience data repositories. Participants emphasized a recent rise in data sharing demands from publishers and funders, leading to a greater need for well-curated datasets in repositories, including thorough descriptions, structured formats, and machine-readability. Furthermore, these repositories have experienced a substantial surge in requests and demands from

various sources. These encompass the storage of vast data volumes, capacity-building efforts for data users, and seamless high-throughput data access for deep learning, especially in the context of emerging foundational models.

In response, data repositories are adapting by offering new access points and quality standards, and some are integrating data with computing resources (cloud or high-performance computing) for end-to-end workflows. Integrating data with computing reduces the need for users to have extensive local computing resources when accessing data and developing data science applications.

6.2.3 Envisioning the needs and actions for geoscience data repositories

The discussion (summarized below) reflects the diverse participants background:

- **Improving data discovery of curated datasets from data repositories.** The expanding volume of geoscience data from various sources makes it difficult for users to find suitable datasets for their needs, particularly in AI application development, which often involves integrating datasets from multiple sources. *To tackle this, data producers and repositories should adopt community-recognized standards with standardized metadata to streamline data discovery.*

- **Providing materials and training to support users to use data correctly.** Capacity building is vital for enhancing the value and broadening the user base of geoscience data. *Activities in this realm include creating computational notebooks and interactive learning materials with real-world geoscience data, which the community has identified as effective methods. Participants also advocated for including comprehensive examples demonstrating efficient access and utilization of curated datasets from repositories.*

- **Adapting to the new and future patterns of data curation and access.** With the rapid development of cloud computing, AI, and other emerging technologies, data repositories need to adapt to the emerging pattern of users' demand on geoscience data. *This adaptation may include improving existing datasets (by working with the owner/producer of the data) to meet users need (e.g., converting data to cloud-native data formats) or developing tools/services/access point that can support diverse users of geoscience data.*

- **Developing a sustainable model to support and maintain data management.** Participants acknowledged the necessity for robust support for data repositories to meet the community's demands for data-intensive applications and quality data management and training. This is a broader community challenge that requires collaborative efforts from various stakeholders to ensure sustainable data management for the scientific community.

This working session marks the beginning of a collaborative community effort to outline key steps for AI-ready geoscience data repositories. The Padlet used in the session will be shared with the broader geoinformatics community through the Council of Data Facilities, a part of Earth Science Information Partners (ESIP), to gather further feedback. We will continue engaging with data users and repositories at conferences and workshops, including the AGU Fall meetings and the ESIP annual meetings to refine our understanding of how geoscience data repositories can adapt to the increasing demand for AI and data-intensive applications. The community's final output will be presented as recommendations for geoscience repositories to enhance their data management and services.

6.3 Session 3: Project-based sustainability for Earth Science Data Infrastructure.

Led by Dr. Arika Virapongse from Middle Path EcoSolutions, this session attracted around 35 participants. The 1-hour and 20-minute working session began with a networking and warm-up activity. Participants were paired with someone they hadn't met before at the workshop and tasked with finding a non-obvious commonality between them. Two rounds of this activity were conducted, followed by a brief share-out to highlight a few examples. The activity received a positive response.

Following the warm-up activity, Virapongse delivered a brief seed presentation available here. The presentation covered key findings from a related research project (Virapongse et al., 2022b, in review) and outlined the session's agenda and breakout prompts.

During the session, participants self-organized into four groups to discuss various aspects of sustainability in their projects. The session slide deck provided a list of guiding questions about their projects, covering topics like project duration, team members, business models, sustainability challenges, lessons learned, and project-specific meanings of sustainability. Some groups discussed these questions individually for each project, while others addressed them collectively in rounds. To foster engaging and inclusive discussions, all group members were encouraged to participate actively in the conversation.

At the session's conclusion, a full-group discussion was held to share interesting insights from the groups' conversations. The topic of project sustainability generated significant interest, with many participants noting its limited depth of exploration despite funders' requests for sustainability statements in proposals. It was acknowledged that project sustainability often requires skills, time, and a shift from an academic to an entrepreneurial model, making it challenging, particularly in terms of securing continued funding or revenue. Some participants shared successful sustainability strategies, such as creative funding through website ads for mineral shows. This highlighted a pressing need within the Earth Science community to continue the discussion on data infrastructure sustainability.

6.4 Session 4: Building upon IoT Projects within the EarthCube Community

The session about IoT environmental sensors and their integration with NSF EarthCube data repositories proved to be both engaging and interactive. Led by Mike Daniels and Daniel Fuka, it began with an overview of the current state of IoT environmental sensing technologies, highlighting recent breakthroughs in microprocessors, sensor technology, and wireless communication. These advancements have significantly reduced costs and improved technology, making cost-accessible environmental sensors accessible to scientists and educators for research and student mentorship. This session stems from the 2019 and 2020 EC All Hands Meetings, where the EarthCube community-initiated working sessions centered around incorporating IoT-based sensor networks and data into early EarthCube Architectures, with some new EarthCube Architectures becoming IoT-based themselves.

After the discussion of IoT applications in environmental sensing, the session transitioned into an interactive phase, featuring a live demonstration of IoT environmental sensors led by one of the participants. This dynamic demonstration showcased IoT sensors' assembly, networking, and live data streaming directly into the NSF-funded CHORDS (Cloud-Hosted Real-time Data Services for the Geosciences) software. Attendees had the unique opportunity to witness the practical application of the concepts from sensing to data workflows to end repository hosting and archiving. All participants actively engaged in a hands-on demonstration assembling IoT environmental sensors and demonstrating how these sensors are constructed and configured for deployment.

The hands-on demonstration included sensor assembly and the network setup, illustrating how sensor data is transported and added to CHORDS real-time data service for the geosciences, which provides an easy-to-use system to acquire, navigate, and distribute real-time data via cloud services.

Attendees were able to observe their personal live data streams coming from the IoT sensors they put together using CHORDS, providing a tangible example of how this technology is leveraged for environmental data collection and analysis.

This hands-on experience allowed attendees to gain a deeper understanding of the IoT sensor deployment process and its integration with data services, reinforcing the workshop's educational objectives. Both the presentation and interactive demonstration ensured a comprehensive and informative workshop experience for all participants, and the session concluded with information and a discussion on how we can encourage funding agencies to support opportunities for IoT environmental sensor projects and collaborative research initiatives.

During this discussion on encouraging uptake by funding agencies, it was suggested that we need to build upon the Birds of a Feather (BoF) community that formed within the EarthCube meetings, and now that the NSF EarthCube 10-year initiative has sunsetted, our "Birds" need to continue to expand our collaborations with researchers, institutions, and organizations that have an interest in the same research area. We should engage colleagues at other confer-

ences/professional organizations (ESIP and AGU were mentioned), and through these workshops, increase visibility and support for cost-accessible IoT sensing. We need to work together to peer review and publish our lower budget preliminary evidence of how it can significantly impact scientific knowledge (NSF), application (USDA, USGS), and society at large (EPA, NASA), emphasize how these impacts are relevant to societal challenges, education, or economic development, that can only be seen with a massive expansion in in-situ sensing that cost-accessibility enables.

During the hands-on segment, discussions within the workshop delved into the current technological and cultural obstacles that researchers have encountered when deploying networks for environmental sensing, giving insight into the challenges and obstacles that still need to be overcome with these technologies.

7 Dissemination of the results of the workshop

This workshop brought geoscientists and cyberinfrastructure experts together to discuss the future of the community, beyond the NSF EarthCube program. This Final Report serves as a summary of the recommendations for future directions made by the community and is made available on FigShare (https://doi.org/10.6084/m9.figshare.23949168).

References

Daniels, M., De La Beaujardière, J.-F., Downs, R. R., Fulker, D., Hills, D. J., Jacobs, G., ... Cramer, C. (2022). *Earthcube sustainability panel report. in earthcube organization materials.* UC San Diego Library Digital Collections. Retrieved from https://library.ucsd.edu/dc/object/bb9634233b doi: 10.6075/J0CR5TJF

EarthCube. (2022). *Funded projects.* Retrieved from https://www.earthcube.org/funded-projects

Jarboe, N., Diggs, S., Downs, R. R., Kinkade, D., Lehnert, K. A., Ramamurthy, M., ... Schreiber, L. (2022). *Cdf sustainability task force report. in earthcube organization materials.* UC San Diego Library Digital Collections. Retrieved from https://library.ucsd.edu/dc/object/bb62553612 doi: 10.6075/J0P55NQJ

Maull, K., & Mayernik, M. (2022). *Earthcube program metrics analysis 2013-2022.* UCAR/NCAR. Retrieved from https://ncar.github.io/earthcube-program-analysis/ doi: 10.5065/VMFJ-QY55

Virapongse, A., Gallagher, J., & Tikoff, B. (in review). *Insights on sustainability of earth science data infrastructure projects.*

Virapongse, A., Gallagher, J., Tikoff, B., Cornillon, P., Koskela, R., Shingledecker, S., ... Hanson, B. (2022a). *Sustainability models for integrated digital earth science. in earthcube organization materials.* UC San Diego

Library Digital Collections. Retrieved from https://library.ucsd.edu/dc/object/bb5606891b doi: 10.6075/J0JH3MBN

Virapongse, A., Gallagher, J., Tikoff, B., Cornillon, P., Koskela, R., Shingledecker, S., ... Hanson, B. (2022b). *Sustainability models for integrated digital earth science. in earthcube organization materials.* UC San Diego Library Digital Collections. Retrieved from https://library.ucsd.edu/dc/object/bb5606891b doi: 10.6075/J0JH3MBN

8 APPENDIX 1: Abstracts

Tuesday, June 27th

Oral Presentations

Brandenberg, Scott: Assimilation of Earth Science and Geotechnical Engineering Data for GeoHazard Assessment

Scientific discovery in the field of geohazards often requires assimilation of datasets from different disciplines. For example, evaluation of earthquake-induced soil liquefaction requires geotechnical data characterizing soil conditions at a site, ground motion data quantifying demand, geospatial data characterizing ground slope, distance to water bodies, etc., and surface geology data characterizing depositional environments. Even when these datasets are publicly available, they are often not integrated in a single computational platform that enables end-to-end workflows. This work focuses on assimilating earth science and geotechnical engineering data using DesignSafe cyberinfrastrcture resources. DesignSafe is supported by the NSF-sponsored Natural Hazards Engineering Research Infrastructure program, and houses a Core Trust Seal certified data repository, applications for interacting with and visualizing data, and provides access to high performance computing resources. Specifically, the presentation will focus on data from two relational databases available to users through the DesignSafe JupyterHub. Using Python scripts, users can query the Next Generation Liquefaction database to obtain geotechnical site investigation data, earthquake ground motions, and observations of liquefaction at a site following an earthquake. Furthermore, users can query a shear wave velocity profile database to obtain geophysical data from various methods including invasive (downhole and cross-hole surveys) and non-invasive (spectral analysis of surface waves, multi-channel analysis of surface waves, horizontal-to-vertical spectral ratio) techniques. The presentation will highlight documented use-case products intended to serve as the basic building blocks of more complicated workflows. The use-case products include Jupyter notebooks that demonstrate how to perform simple queries, extract and plot data, and develop interactive widgets to enable users to create user interfaces.

Curcic, Milan: Challenges and Solutions Toward Efficient Lagrangian Data Analysis for Earth Sciences

Transport problems in Earth sciences, such as ocean garbage tracking, search and rescue, air pollution, and drifting buoys, require Lagrangian data analysis techniques. Unlike Eulerian data which are defined on a fixed system of independent coordinates, Lagrangian data describe the properties of a particle following its position. In the last decade, we have seen tremendous development in the ecosystem of tools and libraries that aid the user with the analysis of gridded (Eulerian) data. However, in the era of cloud-optimized data structures and user-friendly libraries that efficiently handle remote data access, labeled dimensions, and units of measure, working with Lagrangian data remains challenging. In this talk, we will present key technical challenges that end-users face when working with Lagrangian data. The key challenge is that Lagrangian data are not sampled at regular space and time intervals, which makes analyses over fixed space and time windows difficult to implement. Thus, the commonly used grid-based operations from libraries such as NumPy and Xarray do not work if applied on more than one trajectory. To address these challenges, we present a Python library called CloudDrift. Implemented on top of Xarray, CloudDrift aims for a fine balance between ease of use and computational and data-storage performance when handling Lagrangian data. Our talk will summarize our progress and the key technical aspects of CloudDrift's implementation.

Eroglu, Orhan: Project Raijin: Community Geoscience Analysis Tools for Unstructured Grids

Project Raijin has been awarded by the NSF EarthCube program in order to develop sustainable, community-owned tools for the analysis and visualization of unstructured grid model outputs arising from next generation climate and global weather models. The primary development environment for Project Raijin's software tools is the Scientific Python Ecosystem. In particular, the Pangeo packages, Xarray, Dask, and Jupyter provide support for data ingestion and internal representation, scalability, and examples and demonstration, respectively. Two essential goals of Project Raijin are: (1) developing extensible, scalable, open source tools that are capable of operating directly (without regridding to structured grids) on unstructured grids at global storm resolving resolutions in order to support fundamental analysis and visualization methods; and (2) ensuring the long term sustainability of the project by establishing an active, vibrant community of user-contributors that share the ownership of the project and extend our work beyond the scope of this NSF award. This presentation will provide updates about what progress Project Raijin has made to support both of these goals, such as (1) creation of the brand new Python package, UXarray, to provide data analysis and visualization operators on various types of unstructured grids (e.g. MPAS, CAM-SE, E3SM, etc.), and (2) employment of an open development model to encourage community participation in all aspects of the project. We will provide our roadmap for future development and discuss how further community engagement could be possible.

Kumar, Deepak: Urban Heat Island Variability Assessment over Transects for

Climate-Sensitive Sustainable Urban Heat Island Mitigation Approach

The term "urban heat island" (UHI) refers to urbanized regions that are consistently hotter than their rural neighbors. UHI strength depends on factors like urban shape, size, and area, including regional weather conditions. UHI has a significant impact on city residents' lives due to the increased danger of heat-related mortality. Urban regions tend to be warmer than their rural neighbors due to the UHI effect caused by the absorption and retention of heat by man-made materials like concrete and asphalt. Understanding the intensity of urban heat islands enables heat-related health challenges to be better understood. The intended UHI variability assessment can thus be used to locate at-risk regions. Land surface temperatures have been used to estimate surface temperatures in urban and surrounding nonurban areas and to quantify urban heat island intensity. Transects are chosen based on a defined grid of 1km by 1 km, 3 km by 3 km, and 5km by 5km to capture the range of urban surface temperatures. The UHI variability evaluation has been done with the use of a combination of satellite data, ground-based measurements, and modeling techniques. The effects of urban areas on exacerbating heat under present urban conditions are described and quantified in this study. The urban heat island (UHI) variability assessment across transects in the New York City metropolitan area is executed to create long-lasting and efficient mitigation strategies. The UHI variability is then analyzed along transects to identify hotspots and areas with the greatest potential for mitigation. After the assessment, sustainable mitigation strategies, including the use of green infrastructure like green roofs and urban forests, the promotion of energy-efficient buildings, and the implementation of cool pavement and other reflective surfaces, are suggested. Understanding the UHI effect and its impact on the urban environment advances towards creating a more livable and sustainable city; therefore, a proper understanding of the UHI variability over the various transects is a significant step for developing effective and sustainable mitigation strategies. It is endorsed that additional levels of analysis are required as a part of follow-on research to characterize the positive and negative effects of potential mitigation measures. It is also recommended that the analyses of the variability be further studied and quantified to assess the benefits of various parameters for future cool cities.

Mayernik, Matt: Recommendations for accessibility of simulation-based data

It has become a common expectation for researchers to share their data when publishing a journal article or in order to meet sponsor data management requirements. However, many researchers face the challenge of determining "What data to preserve and share?", and "Where to preserve and share that data?" This can be especially challenging for those who run dynamical models, which can produce complex, voluminous data outputs, and who may not have considered what outputs may need to be preserved and shared as part of the project design. This presentation will discuss findings and products from the NSF EarthCube Research Coordination Network project titled "What About Model Data? - Best Practices for Preservation and Replicabil-

ity" (https://modeldatarcn.github.io/). When the primary goal of sharing data is to communicate knowledge, most simulation based research projects only need to preserve and share selected model outputs, along with the full simulation experiment workflow. The rubric was crafted to provide guidance for making decisions on what simulation output to preserve and share in trusted community repositories. This rubric, along with use cases for selected projects, provide scientists with guidance on data accessibility requirements in the planning process of research, allowing for more thoughtful development of data management plans and funding requests. This rubric is being referred to by publishers within journal author guidelines focused on data accessibility.

Working Sessions

Hass, Bridget: Expanding open-source tools for working with aerial remote sensing data in Google Earth Engine

This mini hackathon session will build upon existing Google Earth Engine (GEE) scripts with the goal of developing new workflows and tutorials using the National Ecological Observatory Network (NEON)'s Airborne Observation Platform (AOP) remote sensing data. Potential topics to explore include scaling NEON hyperspectral data to multispectral satellite data, fusing lidar and hyperspectral data, and creating Jupyter notebooks for running geospatial analyses in GEE.

The session will start with a demonstration of a GEE workflow using AOP data, and participants will then be provided several scripts as starting points to build upon. We will provide some potential hacking topic ideas, or participants can work on their own ideas. A useful outcome from this hackathon would be scripts that can be shared with the broader community. These outputs will be converted to tutorials and shared on NEON's website (https://www.neonscience.org/resources/learning-hub/tutorials) as an openly available resource for researchers and educators to use NEON AOP and other remote sensing datasets.

We encourage early career participants to be involved in this session. GEE is a relatively new platform that has rapidly expanded in the last decade. Early career scientists have been increasingly adopting GEE for environmental research, especially as the Big Data revolution has demonstrated the need for cloud computing. We acknowledge that women, people of color, people with disabilities, and other minority groups are poorly represented in the field of environmental data science. Working with large remote sensing datasets typically require experience, computational resources, and data that are not universally available. NEON is working towards shifting this imbalance by providing FAIR (Findable, Accessible, Interoperable, and Reusable) data as well as teaching open-source, reproducible workflows for anyone to use and expand upon. We encourage participants of all backgrounds and at all levels to join this working session, with the goal of developing additional educational resources that will lower the barrier for underrepresented groups to develop skills and contribute to the field of remote sensing and environmental data science.

Requirements: Attendees will need a computer and to register for a Google Earth Engine account. Prior knowledge of JavaScript and/or Python would be helpful but is not required.

Rai, Yuhan (Douglas): Working towards AI-ready Geoscience Data Repositories

The geoscience research community is in a rapid growing phase of AI and other data intensive science. This transition is a great opportunity for geoscience data repositories and facilities to improve its data management and services to serve the evolving user community. There are ongoing activities from the geoscience and machine learning communities to define what AI-ready data means, including NSF Research Coordination Network on FAIR in ML, AI Readiness, and Reproducibility (FARR), Earth Science Information Partners (ESIP), MLCommons, HuggingFace. This working session will leverage these ongoing activities and work with geoscience data repositories and cyberinfrastructure communities to develop a value proposition and recommended practices for data repositories to provide data services for AI and other data-intensive science. This will not only benefit the AI researchers but also improve the data service in general.

The session will be organized as a group writing session using breakout groups to identify the key gaps in how current geoscience data facilities support AI and data-intensive research and produce recommendations to address these gaps. The draft recommendation will be deposited in Zenodo and shared with the geoscience community via FARR, ESIP, Big Data Hub, GO FAIR, and FAIRPoints communication channels to collect feedback. The recommendation will be a living document by incorporating community feedback regularly with proper version control via Zenodo.

To increase the diversity of workforce in geoscience cyberinfrastructure and data facilities, the session will invite early-career researchers and members from underrepresented groups via ESIP, FAIRPoints, and Women in Data Science. To support their participation, we will partner with FARR to provide some partial financial support to reduce the barrier of participation. We also want to work with the workshop organizer to provide hybrid participation method to allow people join the working session remotely.

The session only requires participants to bring their own laptop or notepad to participating in the writing session. Pre-session readings will be shared with the participants in advance in preparation for the session. The session does not require other advanced technologies.

Demos

Elipot, Shane: A Lagrangian analysis workflow using the open source Python library clouddrift

The goal of this demo is to showcase the use of the EarthCube-supported clouddrift Python library that aims at facilitating and accelerating the use of Lagrangian data for climate science. In contrast to Eulerian data that are typically provided on a regular and fixed geographical grid, and thus relatively easy

to analyze conceptually, Lagrangian data are acquired by autonomous platforms that either passively drift with geophysical flows, or actively move with a purpose (think sensor-carrying seals!). Lagrangian data are therefore challenging because of their nonuniform spatial and temporal sampling, as well as their heterogeneity across platforms.

In this example, we tackle the task of estimating the kinetic energy of the near-surface circulation of the global ocean using the high resolution heterogeneous data of the drifting buoys of the NOAA Global Drifter Program, now available through the NOAA Open Data Dissemination cloud storage program. A Jupyter notebook provides the necessary steps to access, select, and process these data to ultimately estimate the spectrum of ocean global kinetic energy continuously as a function of latitude. We will conduct the same analysis with an order of magnitude larger number of trajectories from synthetic particles released in an ocean global circulation model.

Krasnoff, Samuel: VICTOR - A new Cyber-infrastructure for Volcanology Forecasting the impact of active or future volcanic eruption and correctly interpreting the remnants of past eruptions requires access to models of eruptive processes.

The volcano modeling community recognizes a need for more equitable access to robust, verified, validated, and easy-to-use models. To answer this need, we are building VICTOR (Volcanology Infrastructure for Computational Tools and Resources), a new cyberinfrastructure for the volcano modeling community. To date, we have established a steering committee that advises the development team on community needs and best practices. VICTOR is connected with larger, national efforts including CONVERSE and SZ4D's Modelling Collaboratory for Subduction (MCS). We collaborated with the non-profit 2i2c to manage VICTOR's back-end in the form of a JupyterHub in the cloud. We are now developing Jupyter notebooks for the hub, that call existing volcano models such as the lava flow code MOLASSES, the tephra dispersal code Tephra2, and the pyroclastic flow code TITAN2D.

VICTOR will not only provide access to the modeling tool themselves, but also to workflows that utilize these forward models for inversion, benchmarking, and uncertainty quantification. For example, we are in the midst of developing a simple software package for benchmarking pyroclastic density current (PDCs) models against one another. In the future, we hope to expand the tool to add more models, data, and flexibility. Additionally, we are creating tools that integrate sensor data, such as from remote sondes, and utilize convolutional neural networks for debris flow computations and uncertainty analysis using the TITAN2D model. We have already begun a graduate, multi-institutional course on volcanic hazard modeling using VICTOR, with high engagement and collaboration from students. VICTOR is built using open-source tools and hosts primarily open software, emphasizing our commitment to the modernization and accessibility of tools in the volcano science community. Vhub.org, the predecessor, has been migrated to a stub group inside ghub.org as VICTOR grows its library of tools to allow continued access to all resources during the transition.

Wu, Qiusheng: Interactive Geospatial Analysis and Data Visualization with Leafmap

Geospatial data and satellite imagery are increasingly becoming important in various industries, from agriculture to urban planning to environmental management. However, the sheer volume and complexity of these datasets can be overwhelming, making it challenging for individuals and organizations to extract meaningful insights from them. This is where this workshop comes in.

In this workshop, participants will learn how to work with diverse geospatial datasets in the cloud. Through the AWS Open Data Program, petabytes of open geospatial datasets such as Landsat, Sentinel, NAIP, LiDAR, and Maxar Open Data are available for use. We will teach participants how to use these datasets to gain insights into various phenomena, such as land use, vegetation cover, atmospheric conditions, and natural disasters.

To help participants make sense of these datasets, we will introduce them to Leafmap, a Python package that allows users to, search, explore, and visualize geospatial data interactively in a user-friendly and intuitive way. Participants will learn how to use Leafmap to create interactive maps, search for geospatial datasets, visualize Cloud Optimized GeoTIFFs (COG) and SpatioTemporal Asset Catalogs (STAC), and perform basic data analysis tasks in a Jupyter environment.

Throughout the workshop, participants will work with real-world geospatial datasets and will have the opportunity to apply the skills they learn to a range of different use cases. They will also have the opportunity to work in small groups and collaborate with other participants to solve geospatial problems and explore new ideas.

By the end of the workshop, participants will have a solid understanding of how to work with geospatial datasets and how to use Leafmap to gain insights into various phenomena. They will also have developed practical skills that they can apply in their own work, whether they are working in academia, government, or industry.

This workshop is suitable for anyone interested in geospatial data analysis, including researchers, data scientists, GIS professionals, and environmental managers. While some basic knowledge of programming and data analysis is helpful, no prior experience with geospatial data is required. We will provide all necessary datasets and Jupyter Notebook readily available to be used in cloud environments such as MyBinder, Google Colab, and Amazon SageMaker Studio Lab. Participants only need an Internet browser to participate. More information about leafmap can be found at https://leafmap.org.

Posters

Alkaee Taleghan, Samira: Sea Ice Classification by Learning from Label Proportions: How to Classify Sea Ice Type at Pixel-Level using Polygon-Level Ice-Chart Labels?

Authors: Samira Alkaee Taleghan[1], Behzad Vahedi[2], Morteza Karimzadeh[2], Andrew Barrett[3], Walt Meier[3], SiriJodha S Khalsa[3] and Farnoush Banaei-

Kashani[1]
(1) College of Engineering, Design and Computing, University of Colorado Denver, Denver, United States,
(2) University of Colorado Boulder, Department of Geography, Boulder, CO, United States,
(3) National Snow and Ice Data Center (NSIDC), CIRES, University of Colorado Boulder, Boulder, United States

Ice-charts play a crucial role in sea ice monitoring, which is necessary for climate change studies and marine navigation. Traditionally, ice-charts are produced by skilled ice analysts who analyze daily/weekly collected remote sensing imagery (e.g., Sentinel-1 Synthetic Aperture Radar (SAR) data) and annotate the imagery by identifying areas/polygons with a relatively homogeneous distribution of sea ice covers and assigning labels to each polygon. These labels specify overall concentration, the proportion of area covered by up to three main ice types, and flow size. While such expert-generated maps are valuable, manual generation of ice charts is laborious, unscalable, and error-prone, in turn limiting the coverage, recenctly, and accuracy of ice charts. Many previous supervised learning methods have been deployed that leverage expert-generated ice charts as ground-truth labels and train traditional image classification models for automated sea ice classification. However, these methods mainly involve generating pixel-level labels for model training by approximating them from the polygon-level labels available in ice charts, which inevitably limits the accuracy of the corresponding sea ice classification models. In this study, we address this ill-posed problem of training "pixel-level" models from "polygon-level" labels, by utilizing a "learning with label proportion" model for the sea ice type classification. With this approach, we adopt a label proportion generative adversarial network (LLP-GAN) model that directly uses polygon-level ice-chart labels as input for training and learns to generate pixel-level sea ice type predictions as output. By experimenting with different selections of ice charts, we show that the model outperforms existing sea ice classification methods in terms of accuracy while benefiting from more resource-efficient training and reduced training time.

Chakraborty, Sudip: Black Carbon Atmospheric Rivers and Ice Sheet Melting: Machine Learning Approaches.

Increasing temperatures due to global warming and associated climate change impacts are melting the ice sheets in Greenland at an unprecedented rate. Snow darkening by black carbon (BC) aerosols significantly amplifies the greenhouse effect by two times (Hansen & Nazarenko, 2004) and accelerates ice sheet melting (Strong et al., 2009) because they absorb sunlight, warm the surface where they deposit, and darken the snow and ice surface. Aerosol atmospheric rivers (AAR), long and elongated channels of strong wind and extreme mass transport, can transport these aerosols to long distances - often intercontinental. Studies show that BC particles generated over the US and east Asia can often reach Greenland due to AAR activities (Chakraborty et al., 2022; Chakraborty et al.,

2021). Only 20-30 of such activities in a year can transport 40-80% of the total annual transport of BC particles in Greenland. The snow darkening phenomena are very complex and how BC aerosols affect the snow albedo, modulate the radiative properties like surface temperature and radiation, and accelerate the melting process is still unknown. Can snow darkening and associated ice melt be predicted using machine learning?This aims to conduct the causality analysis for ice sheet melting that involves identifying important features that can be used to develop a spatio-temporal machine learning (ML) model to predict ice sheet melting in a sub-seasonal scale (one to four weeks) over Greenland. Currently, we focus on understanding the causality behind ice sheet melting and the factors impacting it by using time series analysis, regression, and the Granger causality method. In the next step, we will use a long short-term memory (LSTM) network model to identify patterns in large data sets, detect correlations, and predict ice sheet melting based on past observational data sets from 2002 -2020 The model will be developed from satellite measurements (e.g MODIS, CERES, AIRS, ICESat 1 and 2, GRACE, SMMR, SSM/I, SSMIS, and NIMBUS-7) using the LSTM method. The predictive model can be used with real-time satellite data (e.g. ICESat 1 and 2 and the data product from MEASURES) and NASA Goddard's information on real-time aerosol transport (Gelaro, 2015) to predict ice sheet melting in real-time. One challenge to address this research question is the various spatial resolution of different satellite data sets ranging between 250 m to 1degree. As a result, all the datasets have been re-gridded using XESMF at 0.25 degree resolution.

Denis, Caroline: GeoSpecify collection management software

Introducing GeoSpecify, an open-source collections management platform for geological museums, designed to manage geological specimens and samples. Based on the Specify software produced by the Specify Collections Consortium, GeoSpecify will support data management and publishing in the disciplines of mineralogy, petrology, and meteorology. The platform will include a comprehensive, role-based access and security system for efficient and scalable user account creation and provisioning. With enhanced intelligence of data forms behavior and validations, accurate data entry is ensured. A powerful "Meta Menu" will provide users with functions for running reports, configuring Carry Forward, inspecting edit history, and more. GeoSpecify will also support community data providers and multilingual functionality. While in the early stages of research and data model conceptualization, GeoSpecify promises to increase productivity with collections data computing, allowing for effective curation, management, and integration of geological collections.

Guo, Tianjing: Towards a Computational Framework for Developing Data-driven Energy Policy

This presentation outlines a vision for a computational infrastructure needed to support the development of data-driven information systems to create a more resilient power infrastructure. The proposed computational framework is a nec-

essary component to address the needs of multiple stakeholders, including individuals, communities, and industry. This presentation highlights the challenges presented when trying to integrate data that capture various aspects of how power demand, consumption, and supply need to be estimated across complex geographical and cultural landscapes.

Nelson, Ellen: StraboSpot: A digital data system for collecting, storing, and sharing geologic data

The StraboSpot (StraboSpot.org) data system was designed through an iterative process with the communities that it serves to facilitate collection, management, integration, and sharing of field and laboratory data in the Geologic Sciences. The goal of StraboSpot is to make these data consistent with FAIR (Findable, Accessible, Interoperable, and Reusable) principles and to integrate multidisciplinary field and laboratory geologic data types into one shared data system. StraboSpot not only provides a shared data repository, but also the tools to collect and manage data and images.

The data system uses the concept of spots, observations that apply over a specified spatial dimension, to nests observations from the regional to microscopic scale and to group data and images as chosen by the user. This approach allows users to connect geologically complex relationships throughout their workflow. The system includes: (1) a mobile application designed to collect measurements, notes, images, samples, and other data while in the field; (2) a web-interface where users can edit, view, and share datasets; (3) a desktop application to manage, store and share laboratory data, including experimental rock deformation data and microstructural images and analyses. These platforms are integrated within the StraboSpot database. Data within StraboSpot may be linked with databases such as Macrostrat, EarthChem, IGSN, and MagIC, increasing the accessibility and interoperability of data stored within StraboSpot and these other systems. Through community input and development, StraboSpot now incorporates structural geology, petrology, sedimentology, and tephra volcanology workflows. The use of controlled vocabularies developed by these communities promotes the standardization of data collection and increases the findability of data. In summary, StraboSpot promotes collaboration among researchers and communities and is an important step in moving the geologic community forward in the digital age.

O'Keefe, Patrick: The Random Hivemind: An Ensemble Deep Learner. A Case Study of Application to Predicting All-Clear SEP Periods.

The application of machine learning and deep learning techniques, including the wide use of non-ensemble, conventional neural networks (CoNN), for predicting various phenomena has become very popular in recent years thanks to the efficiencies and the abilities of these techniques to find relationships in data without human intervention. However, certain CoNN setups may not work on all datasets, especially if the parameters passed to it- including model parameters and hyperparameters, are arguably arbitrary in nature and need to

continuously be updated with the need to retrain the model. This concern can be partially alleviated by employing committees of neural networks that are identical in terms of input features and initialized randomly and "vote" on the decisions made by the committees as a whole. Yet, members of the committee have similar architectures and features passed to them, making it possible for the committee members to "agree" in lockstep. The random hivemind (RH) approach helps to alleviate this concern by having multiple neural network estimators make decisions based on random permutations of features and prescribing a method to determine the weight of the decision of each individual estimator. The effectiveness of RH is demonstrated through experimentation in the predictions of "all-clear" SEP periods by comparing it to that of using both CoNNs and the aforementioned setup of committees identical in input features in this application.

Pandey, Chetraj: Insights into Deep Learning-based Full-disk Solar Flare Prediction with Post hoc Explanation and Evaluation

Authors: Chetraj Pandey, Temitope Adeyeha, Trisha Nandakumar, Rafal Angryk, and Berkay Aydin.

Solar flares are transient space weather events characterized by a sudden brightening of light in the Sun's atmosphere with the enormous release of electromagnetic radiation posing a significant risk to space- and ground-based infrastructures. Therefore, a precise and reliable prediction is essential for mitigating these potential near-Earth impacts. Recent advances in machine learning and deep learning have accelerated the development of data-driven models for solar flare prediction. However, the highly complex data representations learned by these models obscure the transparency and often make it difficult to understand their prediction rationale, which can be problematic in critical applications such as solar flare prediction, where model reliability is crucial. The goal of this study is to extend the coverage area for predictions by being able to issue reliable predictions for near-limb events (beyond +/-70deg in longitude of the solar full-disk). While numerous active region-based flare prediction studies have focused on solar flares occurring in central locations (i.e., within +/-45deg up to +/-70deg in longitude of the disk), we address the problem of issuing a global prediction for full-disk and eventually advancing the prediction efforts by providing forecasts for highly overlooked near-limb events. Furthermore, with post hoc explanations we demonstrate a method to improve the prediction reliability and contribute to the existing cyberinfrastructure of operational forecasting systems. As part of our study, we present a deep-learning model that utilizes hourly full-disk line-of-sight magnetogram images that can predict greater or equal to M-class flares within the subsequent 24-hour window. We apply the Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) attribution method to generate post hoc explanations for our model's prediction and provide empirical findings from qualitative and quantitative evaluations of these explanations. Our evaluation of the model's explanations demonstrates that it is capable of identifying and utilizing features associated with active

regions in both central and near-limb locations in full-disk magnetograms to produce accurate predictions. Notably, the model can recognize the shape and texture-based properties of active regions, even in near-limb regions, which is a critical capability with significant implications for operational flare forecasting systems.

Porter, Brent: Model Evaluation and Execution Platform: System architecture for model management and storage to support integrated planning and interoperability at scale

The model evaluation and execution platform (MEEP) is a multi-functional set of services to support integration between data and models, and streamline common tasks needed to incorporate results into decision support processes. MEEP is in development with initial model management services in an operational environment to support statewide flood planning in Texas.

MEEP is designed to streamline connections across sets of services needed for common modeling tasks, including:

- Core Storage: General data ingest to load heterogeneous data collections from long tail, or small data, up to very large data objects.Model Evaluation: Standardized virtual environments using a container orchestration system for sharing and peer review of model versions between users to support technical review and versioning of simulation models.

- Model Execution: An HPC allocation and environment to support simulation model execution at scale on compute services adjacent to data storage. Integration Platform: A custom-built operational model integration suite, (MINT) brings together these components and allows for the execution, configuration, discovery and analysis of models and data through a set of web-based tools.

The initial proof of concept architecture and application for MEEP provides an end-to-end roadmap for generalizable architecture to support integrated modeling with features for data ingest, discovery, review, versioning, workflow engineering, and provenance tracking. MEEP facilitates collaborative applications and services with local, state and federal partners. Particular attention is given to the APIs and middleware as a means of prioritizing interoperability. MEEP is in active development with operational components currently in use for flood planning and disaster information data services."

Virapongse, Arika: Sustainability of Earth Science data infrastructure projects

Cyberinfrastructure is an essential part of how science is done today, but sustaining the projects that provide these services is a challenge. In this study funded by the Council of Funded Projects of the NSF EarthCube, we examined eleven long-term data infrastructure projects, most focused on the Earth Sciences, to understand the characteristics that contributed to their sustainability. Among our sample group, we noted the existence of three different

types of project groupings: Database, Framework, and Middleware. Database projects aim to bring together data and data resources for use. Middleware projects seek to develop software and technology. Framework projects focus on developing best practices.

The conflicting expectations, limitations, and needs of academia and cyberinfrastructure development presented challenges for all project types. Although none of the studied projects began with a formal governance model, each project adopted a model over time. Most of the efforts started as funded research projects, and nearly all became organizations in order to become sustainable. Projects were often funded for short time scales, but had the long-term burden of sustaining and supporting open science, interoperability, and community building-activities that are difficult to fund directly. This transition from "project" to "organization" was challenging for most efforts, and specifically in regards to leadership change and funding issues.

Some common approaches to sustainability were identified within each project grouping. Framework and Database projects both rely heavily on the commitment to, and contribution from, a disciplinary community. Framework projects often used bottom-up governance approaches to maintain the active participation and interest of their community. Database projects succeeded when they were able to position themselves as part of the core workflow for disciplinary-specific scientific research. Middleware projects borrowed heavily from sustainability models used by software companies, while maintaining strong scientific partnerships. Cyberinfrastructure for science requires considerable resources to develop and sustain itself, and much of these resources are provided through in-kind support from academics, researchers, and their institutes. It is imperative that more work is done to find appropriate models that help sustain key data infrastructure for Earth Science over the long-term.

Tuesday, June 28th

Oral Presentations

Clyne, John: Update on Project Pythia: A Community Resource for Geoscientific Python Education

Project Pythia launched in 2020 with NSF EarthCube support to be the educational arm of Pangeo and serve as a community training resource for Python-based geoscientific computing. Pythia has a strong focus on the Pangeo Stack of packages (Xarray, Dask, and Jupyter). Pythia has two core goals: (1) reducing knowledge barriers by developing open, interactive, web-accessible learning resources built on public, cloud-hosted datasets that "just work" for users; and (2) growing an inclusive Open Science community around this content.

To date, Pythia has developed two primary educational resources: Pythia Foundations, and Pythia Cookbooks - housed on open GitHub repositories and served via our website. Foundations is a geoscience-flavored introduction to the essential tools in the Scientific Python Ecosystem and Pangeo stack (e.g., JupyterLab, NumPy, Matplotlib, Pandas, Cartopy, Xarray, Dask), plus Python

environment management tools (conda), basics of version control (git), and effective use of GitHub as an open source communication platform. Cookbooks are crowd-sourced collections of advanced, domain-specific tutorials and exemplar workflows (or recipes) that build upon Foundations with explicit links to necessary background knowledge.

As we conclude the third and final year of funding for Project Pythia, this talk will provide a synopsis of Pythia's current resources, share our experiences with growing a sustainable community of contributors, and present our plans for the future.

Deauna, Josephine Dianne: Improved grid-aware operations for ocean model analysis: expanding metrics functionality in the xgcm Python package

Ocean models of varying resolutions are powerful tools for analyzing and predicting ocean states over historical and future time periods. These models divide the ocean into 3-dimensional cubes or grid cells, where averages of ocean variables per cell are calculated by integrating partial differential equations forwards through time. Depending on the configuration, model grids of scalar (e.g., temperature) and vector (e.g., velocity) quantities can be staggered with respect to each other within a grid cell. Model metrics define the relationships among those different positions (e.g., the distance along the x-axis between two temperature vs two velocity points). These are essential in post processing analysis, for example, when computing physical quantities for curvilinear models with non-uniform distances between cells. The xgcm Python package was developed to enable easy application of operations such as averaging, integration, and differentiation (among others) across different model configurations in a convenient and highly efficient manner, by maximizing the utility of model metrics when doing calculations. This talk will focus on updates developed for xgcm's handling of grid cell geometries in GCMs, more specifically on assigning model metrics, guessing them when necessary by interpolating across different dimensions, and selecting the appropriate metric for a given operation.

McHenry, Kenton: Ongoing Impacts From the EarthCube Technical Efforts

To address the challenges surrounding data and tools within the geoscience community over its years the EarthCube effort identified and settled on two approaches that not only addressed many of these challenges but were also sustainable past the program. With regards to data, conventions on top of schema.org, called science-on-schema, were adopted and support for its usage was put in place so that geoscience data repositories would support it, allowing for datasets to be indexed from across community repositories and enhancing discoverability. With regards to software the notion of peer reviewed notebooks was leveraged to motivate the community itself to document and put out their software in a usable manner for others to use, enhancing reuse as well as the impact of developed scientific software. These two approaches adopted by the geoscience community within EarthCube have since begun to move on to be adopted by other fields in order to address these similar challenges. In this

talk we will describe how these EarthCube activities are now being leveraged in areas such as ecological forecasting, geochemical science, deep ocean science, amongst the emerging research software engineering community in the U.S., as well as within publication with publishers such as the American Geophysical Union and Wiley.

Moon, Seulgi: Center for Land-Surface Hazards (CLaSH): Bridging Earth Data and Cascading Hazard Processes

Land-surface hazards, such as landsliding and river flooding, have an enormous impact on humans because they occur frequently in many environments around the world. As these geohazards magnify due to climate change and human activity, there is increased urgency to understand and predict their future effect. Moreover, the complex interactions of slope and river erosion processes often catalyze "cascading hazards", where initial events trigger subsequent ones that magnify hazards for years to decades. Research on cascading hazards is a frontier prime for major advancement, but accurate assessment of these threats requires the integration of a wide range of different types of data by interdisciplinary research teams. The NSF-sponsored Center for Land-Surface Hazards (CLaSH) Catalyst project is developing a shared vision within the scientific community around innovation in hazards research and education. Within the CLaSH vision, we aim to engage expertise from geoscience, engineering, and climate-related fields, as well as, to build strong relationships with existing NSF centers/facilities that provide data collection and data curation. A proposed partnership between DesignSafe and CLaSH Catalyst project provides one such important opportunity to engage in cross-disciplinary efforts to develop open-source data and modeling resources that broadly serve the geoscience community and foster innovation in hazard science. A proposed pilot database will be focused on landslide susceptibility with a "critical zone" framework. The physical and chemical processes in the "critical zone" break down competent bedrock into transportable materials, which is crucial to assess landslide hazards. We propose to adopt this critical zone framework in constructing the array of different data constituents that include both physical and chemical ground-based measurements in landslide-prone areas.

Quinn, Daven: Crossing the "collaborative barrier" for Macrostrat and digital crustal research

Macrostrat (macrostrat.org) is a platform that integrates stratigraphic columns and geologic maps into a digital description of the Earth's crust. This data system has become a widely used research tool that describes the Earth's geological record in space and time. Its global, harmonized geologic map and associated stratigraphic, fossil, and age information are in wide public use for education and outreach. For instance, the Rockd mobile app (rockd.org), which provides exploration and citizen-science capabilities atop Macrostrat's data holdings, has recently surpassed 98,000 users, with hundreds of contributed of outcrop "checkins" each month.

Macrostrat's continued impact relies on its continued ability to expand its core data holdings. Increasing resolution and spatial coverage will allow Macrostrat's digital approach to be applied to a wider range of study areas and scientific problems, while also increasing its value as a contextual data resource. Currently, all stratigraphic and map data ingestion is done by a single lab group, which is now limited in both capacity and regional expertise to expand data holdings beyond North America. This limitation can be overcome by building tools for collaborative management of the archive by a wider set of geologists. This expansion requires both new software infrastructure and establishing productive integrations with individual researchers and geologic surveys positioned to contribute to such a system.

The current iteration of Macrostrat was constructed with support through EarthCube over the last decade, and the development of a broader ecosystem for crustal data management will likewise be reliant on NSF funding. However, longer term, the flourishing of this research approach is more limited by the structure of the geologic research community. Currently, few researchers possess the combination of geological training and technical skills required to build effective software infrastructure. Relatedly, collaborative maintenance of research software has not yet become an accepted mode of geologic research, leading to unproductive "silos" of work.

To address these adverse structures, we are working to establish the DigitalCrust organization (to be launched in Fall 2023 at digitalcrust.org and with a GSA workshop). This consortium, which takes cues from efforts such as Software Underground, will seek to build strong collaborative ties between geoinformatics researchers across organizations and to situate maintenance of shared software libraries as increments of research progress. It will also train geologists in the basic practices of collaborative software development. If successful, this "low-level" approach will help enhance community buy-in towards maintaining software infrastructure for geological research. This will ultimately benefit of Macrostrat and many other related efforts, contributing to eventual transdisciplinary digital models of Earth's crustal structure and evolution.

Working Sessions

Fuka, Daniel: Building Upon the IoT Projects within the EarthCube Community

Session Leaders: Daniel Fuka, Mike Daniels, Agbeli Ameko, Keith Maul, Mike Dye, Je'aime Powell, Ruth Duerr

Activities: This will be an open-room collaboration-making through demos session, with outcomes focused on forming new multi-community cross-cultural collaborative research projects.

Outcome/Product: New Collaborations among researchers new to the community and existing EarthCube Alumni.

Outcome Sharing: A short report detailing new project proposal abstracts and impact statements will be created during the session.

Materials/technology: Participants and leaders will supply all materials and technologies.

Early-career Involvement: Early career participants will directly interact with the EarthCube Alumni and lead in developing project proposal abstracts and associated impact statements.

Underrepresented group involvement: IoT projects are especially critical to underrepresented groups in the Geoscience Cyberinfrastructure research community, specifically as they provide affordable research quality sensor data gathering, workflows, and data archiving and analysis to researchers with limited or non-existent research budgets. Only through the use of off-the-shelf IoT components can many communities afford to monitor their geoscience surroundings.

Session Overview: The EarthCube community brought forward many architectures that ended up supporting the rapid advances and decreasing costs in microprocessors, sensing technology, and wireless communication, which enables scientists to deploy environmental sensing technologies with lesser budget requirements. These IoT-based sensing capabilities allow environmental processes to be continuously monitored in habitats ranging from very remote to urban, providing information in unprecedented temporal and spatial resolution. At the 2019 and 2020 EC Allhands Meetings, we expanded working sessions centered around incorporating IoT-based sensor networks and data into early EarthCube Architectures. Many of the groups have taken these capabilities and expanded them beyond their initial EC Funded projects. The focus of this working session is not to cover data acquisition systems that already have workflows and recommended standards and specifications but rather to bring together the cluster of scientists who have continued these citizen science IoT-based data acquisition technologies that have and continue to Build Upon the EarthCube and Partnering Geoscience Cyberinfrastructures which enabled uniqueness as the persons who build and deploy their sensor suites. We plan to expand the loosely coupled community and expand on this working session during this EC Geoscience and Cyberinfrastructure Workshop, an excellent opportunity for the EarthCube Legacy to expand its reaches deeper into field-based community research science campaigns throughout the geosciences.

Virapongse, Arika: What's next?: Project-based sustainability for Earth Science data infrastructure

Session Leaders: Arika Virapongse and Julie Newman

Session Overview: A welcome by group leaders. A short introduction by Virapongse on the concept of sustainability and the bullet point findings from the study Virapongse et al (in review) that was funded by the Council of Funded Projects (if not given as a presentation).We'll also provide some tips for how to be inclusive of others and mindful of power dynamics that often occur in group gatherings. We'll ask people to pair with the person next to them and introduce themselves. Then, each person will introduce their partner.

Key Session Activities: We will break into groups of 6 who will share their thoughts/experiences on making their individual projects sustainable. As a guide, we'll use some of the interview questions that were used in the study Virapongse et al. With the help of a facilitator, each group will select a time

keeper, scribe, and presenter (early career preferred) before beginning the discussions.

Each breakout group will share ideas with bigger group.

In a facilitated discussion, group will evaluate if there is a follow on activity that should be conducted.

Key Session Outcome/Product: We will produce a list of ideas for how to achieve sustainability for the project and for the geoscience data community overall.We will potentially produce/propose possible networks for future collaborative work.

How to share outcomes: We will share the outcomes of this work to the larger community via the EarthCube listserv.

Early-career participants: The introduction is designed to help early career participants feel more comfortable by having a personal interaction with another participant. They will be encouraged to act as presenters for the breakout groups. One session leader is early-career.

Underrepresented groups: At the start of the session, organizers will conduct a small exercise to help participants be mindful of positions of power they hold within the group. During the group interactions, facilitators will encourage inclusive behavior, such as ensuring that each person gets to speak equally, recognizing interruptions, and empowering less confident participants. Both session leaders are from underrepresented groups.

Demos

Ameko, Agbeli: Rural Community Engagement Through OpenIoTwx and EarthCube's CHORDS Platform

Agbeli Ameko, Keith Maull (National Center for Atmospheric Research), Mike Daniels, Mike Dye (Ronin Institute), Daniel Fuka (Virginia Tech), Melissa Waters (Pueblo Community College Southwest), Cherie Brungardt (Northeastern Junior College), Linda Hayden (Elizabeth City State University), Francis Tuluri, Remata Reddy (Jackson State University)

Workforce diversity in the geosciences can be enhanced by embracing not only science expertise but also engineering, software, data wrangling and other development and support areas. Furthermore, bringing climate data sampling to local communities can pique the interest of a future workforce interested in climate, weather and more broadly science, engineering and technology. Fortunately, we are in the midst of a revolution of ubiquitous sensor data which is democratizing data access to an unprecedented level. NCAR's OpenIoTwx platform (https://ncar.github.io/openiotwx/), consisting of 3D printed parts as well as very inexpensive electronic components and sensors, has tremendous potential to bring local weather (and other community) data to traditionally underserved communities. By embracing the diverse skills needed to construct these IoTwx platforms, placing them at or near libraries or community colleges in rural communities and making the local weather data easily accessible through EarthCube's Cloud-Hosted Real-time Data Services for the Geosciences (CHORDS, see http://chordsrt.com) project , we are stimulating interest in

student and citizen science, the geosciences and data-driven community decision making across a wide range of community partners and stakeholders. We envision a whole new set of data products coming into maturity through our aims of building simplified, state-of-the-art, robust, quality-controlled and "born connected" measurements more cheaply and more broadly than ever before.

Threats to rural communities are particularly large given increasing climate change impacts, and with both historical and real-time data often in short supply, decision-making, modeling and planning are rendered ineffective during short-term crises and trend analysis for medium term planning lacks statistical certainty. Through the collection and access of these data, an IoTwx station positioned in a local community could enhance understanding of severe weather phenomena and spark interest in atmospheric science and STEM careers. As part of a small NCAR Diversity grant, we have first partnered with rural Colorado Community Colleges and are working with STEM educators to engage 2YC students. In the future, we are working with colleagues to expand this effort to include Historically Black Colleges and Universities (HBCUs) to grow a broader Open IoT STEM Learning Communities program. If our abstract is selected, we will set up a demo of a prototype version of our modular weather station designed with funding from this modest project.

Jarboe, Nicholas: Using the Modular FIESTA Software Stack for the Quick Stand-Up of FAIR Data Repositories for Drilling Core Multi-Sensor Track Data (CDR) and 40Ar/39Ar Age Data (KARAR)

The Framework of Integrated Earth Science and Technology Applications (FIESTA, https://earthref.org/FIESTA) is a containerized set of services designed for reuse which enables the quick stand-up of sample-based geoscience subdomain data repositories. We have been funded by NSF under the EarthCube project to use FIESTA, which we developed to support the MagIC (https://earthref.org/MagIC) rock, geo, and paleomagnetic data repository, to create data repositories for the core multi-sensor track data community and the 40Ar/39Ar rock dating community. Using the FIESTA software system and data models created in collaboration with the scientific communities that will use the data repositories, we have created beta versions of these repositories with a small number of example datasets. These beta sites include most of the features included with FIESTA such as ORCID iD for identity authentication and secure login, quick text searches using Elasticsearch, complex searches using ranges over multiple data columns, the ability to combine search results from multiple datasets into a single file download, schema.org/JSON-LD headers for every dataset for indexing by EarthCube's GeoCODES and Google Dataset Search, a data DOI minted for each data publication, data validation based on the data model, customizable website homepage layouts showing the most recent data publications and community events, and a private workspace where researchers can upload data before publication with the option to share with colleagues, journal editors, or reviewers. These beta sites are being used to solicit feedback from scientists to make sure the repositories meet their needs and

to enable quick, iterative improvements to the repositories.

Zhang, Jiyin: OpenMindat Data API: the Automated Gateway to Two Decades of Crowd-Sourcing Mineral Data

Mindat.org, known as one of the largest and most comprehensive mineral information databases, has accumulated extensive crowdsourced data over the past two decades. To facilitate machine access to this valuable resource and promote collaboration between mineralogy and data science, with support from the NSF EarthCube program, we have developed the OpenMindat Data API. This functional tool simplifies data retrieval by providing essential filtering and selection capabilities, enabling users to quickly query and download structured mineral data. The API covers various aspects of mineral information, including mineral names, chemical formulas, classification hierarchies, localities, occurrences, and more. In the past few months, we wrote up a series of documentation about the API (see the tutorial at: https://www.mindat.org/a/how_to_get_my_mindat_api_key) and collaborated with several geologists to develop applications with datasets retrieved from the API. In this presentation, we will demonstrate several use cases that showcase the convenience and practical usefulness of the OpenMindat Data API in advancing interdisciplinary research and applications. We welcome any interested users to try the API, and we will be happy to have interactions and answer questions. This work is supported by the National Science Foundation (#2126315).

Posters

Agarwal, Khushboo: Modeling electric grid vulnerability induced by natural events using Machine Learning and Geospatial analysis

The growing frequency of weather-induced power outages in recent decades has put the electric grid infrastructure of the United States at risk. Natural hazards, like hurricanes, floods, heat waves, and winter storms, can cause millions of dollars of loss to the grid infrastructure. Any damage to the electric grid can further impact other critical infrastructure, like water distribution and transportation. Past instances show that these events have more impact on low-income communities. Therefore, modeling the grid vulnerability to weather extremes is vital to protect these communities. This research involves the creation of a multi-step modeling method to predict the spatial extent and number of electric power outages for a case study in the Rio Grande Valley region of Texas. The study focuses on the impact of hydrologic flood events on the power grid using a step-wise workflow that scales geospatial analyses and applies a machine learning approach to inform prevention, mitigation, and restoration strategies.The initial analysis generates a flood inundation model using the Height Above Nearest Drainage (HAND) method. This python workflow uses the Stampede2 supercomputer to produce the HAND flood extent for one input DEM tile in 16 seconds and was made scalable for 135,000 DEM tiles in the Rio Grande valley. The implementation presents ways to scale up and time-bound such hydrologic models on high performance computing systems. Combining

the HAND data product with Precipitation Frequency Estimate generated a Flood Vulnerability Raster (FVR) to provide the base dataset for subsequent steps in the analysis. The areas most prone to outages are identified using a spatial power outage model that combines information from the grid infrastructure maps, FVR, Social-Vulnerability Index (SVI) data, and the Weather Events dataset. The same set of input datasets with a total of 47 features for each of the 17 counties in the Rio Grande, along with their historical power outage data, is used to train a Random Forest model to predict the power outage expectation for a county. The Random Forest model performs well with a low normalized RMSE of 11.8%. Also, an analytical model for future outage prediction is developed based on linear regression. The simplified power-outage analytical model is created by using feature selection and utilizes only four important variables for an R-squared of 0.88. Furthermore, this research discusses possible practices that can improve power system resilience, such as deploying microgrids, expansion of transmission capacity and grid hardening. Research results show promise for use by urban planners, operators and decision-makers that make decisions related to resource allocation, critical infrastructure protection, investments, and manage emergency preparedness.

Emile-Geay, Julien: Pandas and the geosciences: a 4.5 billion year story

Timeseries analysis underlies many fields of science and engineering, including many of the geosciences. The paleogeosciences (paleoclimatology, paleoceanography, paleontology, paleoecology, archeology) present unique challenges to timeseries analysis: the time axis is often unevenly-spaced, time is highly uncertain, timescales vary from days to billions of years, time is represented as positive towards the past ("age") with an origin point that is often defined based on the dating method (e.g., A.D. 1950 for radiocarbon measurements, A.D. 2000 for U/Th). These challenges have made it difficult for the paleoclimate community to use standard libraries. For instance, standard NumPy datetime64 objects are the bedrock upon which many other time-aware libraries, such as Pandas, are built. Up to this year, however, NumPy datetime64 objects were only implemented as nanosecond resolution, limiting the time span that can be represented using a 64-bit integer to a geologically inconsequential 584 years. This cut off important geoscientific domains from using pandas.

Here we describe the implementation of a non-nanosecond dtype for datetime64 objects in Pandas, allowing resolutions as coarse as 1s (and therefore a timespan of a few billion years). We illustrate its use in the Pyleoclim library, designed for the analysis and visualization of paleoscientific timeseries, which cover timescales of days to billions of years. We show that this enhancement allows for much more robust capabilities, and enhances interoperability with other popular libraries, such as Xarray. Although the Pandas extension and incorporation into Pyleoclim represents a major stepping stone to allow scientists in these domains to make use of more open science code, work remains for interoperability with other open source libraries such as Matplotlib, Seaborn, Scikit-learn, and Scipy. These libraries will need to update for non-nanosecond

to fully unleash the power of the Python open source ecosystem for the paleogeosciences. Furthermore, our work highlights the importance of well-defined calendars for conversion between dates. For example, the standard assumption of 24 hours per day does not apply in the past and model simulations often do not use the Gregorian calendar.

Fletcher, Lydia: Improving Traceability Throughout the Data Lifecycle: the DOLCE Approach to Provenance

The Texas Advanced Computing Center (TACC) uses a whole lifecycle approach to data management called the Digital Object LifeCycle Ecosystem (DOLCE). The overall goal of DOLCE is to create policies and services that enable TACC to support accessibility to and discovery of digital objects throughout the phases of their lifecycle including generation, processing, description, analysis, storage, and sharing. The ultimate goal of DOLCE is to produce data that aligns with the FAIR Principles of findability, accessibility, interoperability, and reusability. A key aspect of this data curation process is using data provenance to improve reusability of data by tracking the processes used to create, gather, transform, and analyze digital objects. In this poster, we present our goals for improving data provenance using GIS data as a use case. We will demonstrate how we utilize robust metadata to capture important data processing steps. We will also explain how provenance ties into our development of a data catalog that promotes long-term preservation and reusability.

Landers, Jordan: PaleoBooks: A Library of JupyterBooks for Paleoclimate Research

Data-Model comparison is a vital part of climate research, but wrangling the often nonstandard observational data and large model output files involved requires researchers to transition from analysis formats like spreadsheets to a programming language like Python. This transition can be unwieldy, so digestible, topically relevant resources are essential to facilitate it. Here we report on a library of JupyterBooks, called PaleoBooks, that provides guidance and templates for addressing common challenges in the more data-intensive aspects of paleoclimatology, particularly data-model comparison. Each themed PaleoBook includes technical Python lifehacks and scientifically-oriented research workflows for accessing, tidying, and navigating relevant data, performing topic-specific calculations, and thoughtfully visualizing observations and model output. These JupyterBooks are available from LinkedEarth and come complete with configuration files that make it easy to build the right coding environment in the cloud or locally for interactive exploration and repurposing. We showcase the key features of the PaleoBooks library and demonstrate its usefulness in streamlining paleoclimate data-model comparison.

McKay, Nicholas: A Paleoclimate Reconstruction Storehouse (PReSto) to integrate paleoclimate data, reconstruction workflows and visualization

Paleoclimate reconstructions are among the most widely used scientific prod-

ucts from the paleoclimate community. A single chart like the "Hockey Stick" temperature reconstruction of the past 1000 years synthesizes what is known about past climate variations in a form that is easily digestible within and beyond the geosciences. However, such reconstructions are infrequently updated and commonly lag many years behind the latest data and methods and the lack of a central clearinghouse also makes them difficult to find. Finally, paleoclimate reconstructions involve potentially many subjective choices that have not been exhaustively explored, and are opaque to most users.

Here we present the Paleoclimate Reconstruction "Storehouse" (PReSto). PReSto integrates workflows that (1) Draw from the most up-to-date, curated paleoclimate datasets and compilations;(2) Apply an array of published methods to produce continuously-updated reconstructions; and(3) Provide effective access in a responsive web front end, allowing users to easily visualize, download and compare published reconstructions.

Data enter PReSto workflows as metadata-rich Linked PaleoData (LiPD). The LiPDverse (https://lipdverse.org) serves as a versioned, queryable data service for paleoclimate datasets and compilations that supply data to PReSto reconstruction algorithms. We are piloting PReSto with reconstruction algorithms that automatically update reconstructions with new and updated data for key use cases that span different timescales over the past 12,000 years, and different methodologies, ranging from simple compositing to data assimilation.

To standardize PReSto workflows, we've developed a method for containerizing these algorithms to accept input data from the LiPDverse and a set of standardized parameters in JSON. This standardization makes the methods modular, which will ultimately allow users to test the impacts of changes in methodology in addition to input data and parameters. PReSto provides access to these containers with a web interface, where computing is performed on a remote server and users are given access to the resulting data and graphical outputs by direct download. More sophisticated users can also download the containers themselves for increased flexibility.

We have also developed a python framework that takes the standardized netCDF output of the reconstruction algorithms and creates a collection of webpages to visualize results. These webpages allow users to browse a large collection of maps and time series to better understand paleoclimate reconstructions in an intuitive framework. The website also provides descriptions of methodologies, links to papers, and access to reconstruction data and code, so users are well equipped to pursue more specific research questions themselves. The website will also be updated with newer versions of reconstructions, so users can explore how the results evolve as the proxy network expands.

Samykannu, Venkadesh: Spatio-temporal trend analysis of satellite-based CHIRPS precipitation (1981-2017) and significance of crop production in selected crops over Tamil Nadu, India

Agriculture in Tamil Nadu is highly dependent on rainfall and its distribution. Understanding the Spatio-temporal variability of precipitation on crop

production is an immediate necessity in the present climate change scenarios. Tamil Nadu is situated in a very crucial place in the Indian sub-continent and in recent years it is experiencing droughts and extreme rainfall events at irregular intervals which are making crop growth uncertain. Heavy urbanization in recent decades may aggravate this in the coming years. The objective of this study was to investigate the Spatio-temporal pattern of annual rainfall using the Innovative trend analysis (ITA) method for long-term time series of satellite-derived precipitation from Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) and district-wise yearly production data for the crops viz., Paddy, Groundnut, Cotton and Onion from International Crops Research Institute for the Semi-Arid Tropics's (ICRISAT) District Level Database were collected during the period from 1981 to 2017 (37 years) for 13 districts over Tamil Nadu. The results of long-term trends obtained from the ITA method detected significant increasing trends at the 99% confidence level in annual rainfall in almost all the districts of Tamil Nadu. High rainfall from the long-term rainfall pattern was one of the major environmental factors for high productivity. It shows a strong correlation between spatial and temporal variations between precipitation and crop production in Tamil Nadu. The outcomes revealed that the ITA method is sensitive to detecting trends and helpful for long-term sustainable crop production in Tamil Nadu.

Sun, Alex: Building an AI/ML framework for flood inundation prediction

Each year flooding causes substantial economic losses and affects millions of people globally. Climate change has further exacerbated the frequency/magnitude of floods. Thus, a strong need exists for flood early warning systems and real-time flood risk management capability at all scales, calling for faster and more accurate flood inundation models (FIMs). Under a coastal digital twin project, we have been developing a hybrid AI/ML FIM by leveraging the strengths of full shallow water equation (SWE) solvers for high-resolution flood mapping and the flexibility/efficiency of deep learning for nowcasting. A unique challenge in building such earth system digital twins is related to the acquisition and processing of large volumes of earth observation data and streamlining machine learning operations (ML-Ops). We have leveraged a large number of Python libraries originated from the scientific computing community. In this presentation, I will demonstrate an urban flooding ML-Ops pipeline for the Houston TX metro area. Specifically, training data were generated by an SWE solver using high-resolution digital elevation model (3-10m) and 2-min nexrad rainfall data corresponding to a number of storm events. A divide-and-conquer strategy was used to perform ML on small randomly sampled data cubes, instead of at the basin level. Preliminary results suggest such a strategy led to reasonable results, potentially enabling the ML to scale up to much larger watersheds.

Valentine, David: The Quirks of Science on Schema: A Retrospective on Geocodes, and a Path Forward as Decoder

Science On Schema is flexible by design providing many paths to imple-

mentation. NSF Earthcube Geocodes harvested Science On Schema JSONLD information from 27 Council of Data Facilities Projects. The data was ingested using the GleanerIO software stack. While several communities use the GleanerIO stack, OceanInfoHub, Polder, and InternetOfWater, Geocodes is unique in that it is not a single community focused on a set of standard representations. Geocodes sought to ingest a many communities SOS, meaning that diversity had to be addressed in order to render the information and speed retrieval. To improve search performance, we implemented a materialized information view, and a javascript UI was left to deal with many quirks of implementations. JSONLD allows for diverse implementations, for example how identifiers are defined. The 'identifiers' property is used for communicating DOI's, and/or local identifiers, and there should be one or more identifiers. But, identifiers can be a simple string `identifier`: `https://doi.org/10.1234/1234567890` or an object `identifier`: `@type`: `PropertyValue`, `propertyID`: `https://registry.identifiers.org/registry/doi`, `value`: `doi:10.1234/1234567890`,`url`: `https://doi.org/10.1234/1234567890` or an array of objects. In such cases, Geocodes UI rendered diverse representations. Where quirks were a frequent implementation issue, we coded workarounds. For example, bad/improper JSONLD contexts prevented data ingestion, and are a common issue. We implemented fixes to allow for the ingestion of the data. Still incorrect implementations were an issue. One example was unique id's (uid). A document may contain many uid's (json property: @id), and each should be unique. If multiple 'objects' use the same ui, then when converted into a graph representation, objects are merged. A single 'object' is retrieved from the graph, with two 'types.' The context that one of the objects could be a property of the other is lost, causing discoverability and use issues. In such cases, we asked data sources to address these changes.

For DeCoder, we seek to better communicate with communities what is in (or missing) from their schema.org information. Q/A tools help with the data ingestion process: what pages do not have JSONLD, what JSONLD did not convert to a graph, what types are utilized by a data source, etc. We are validating the quality of the information using SHACL 'shapes.'

For DeCoder, we plan on working on the linkages between community tools and available data. We will focus on three communities: low-temperature geochemistry, ecological forecasting, deep ocean observing. As part of Decoder, we plan to standardize and enhance the information model and load a single model into the graph. This should improve discoverability, and allow for tools to data linkages to be achieved across diverse community representation.

Zhu, Feng: cfr: a Python package for climate field reconstruction

Climate field reconstruction (CFR) is the emerging approach to study the spatiotemporal climate history of the past and perform out-of-sample validation of the climate models. Its whole workflow can, however, be complicated, time-consuming, and error-prone, which usually involves preprocessing of the proxy records, climate model simulations, and instrumental observations, application

of the reconstruction methods, and analysis and visualization of the reconstruction results. cfr is an open-source and object-oriented Python package that aims to make the CFR workflow easy to understand and conduct, saving climatologists from technical details and facilitating efficient and reproducible researches. It provides user-friendly utilities for common CFR tasks such as proxy and climate data analysis and visualization, proxy system modeling, and modularized workflows for multiple reconstruction methods, enabling inter-methodological comparisons within a same framework. As an illustration, we present two cfr-driven reconstruction experiments taking the last millennium reanalysis (LMR) paleoclimate data assimilation (PDA) approach and the Graphical Expectation-Maximization (GraphEM) algorithm, respectively.