1. **Project Details**

Name of project: Cloud-based Open Science Machine Learning Tutorials for Earth Science
Project lead and contact details:

- Lead: Yuhan (Douglas) Rao (yrao5@ncsu.edu), NCICS/CISESS/NCSU, affiliated with NOAA NCEI
- Co-lead: Christopher Slocum (christopher.slocum@noaa.gov), NOAA/NESDIS/STAR

Proposed start and end date: March - October 2021
Budget requested: $10,000

2. **Project Outline**

2.1 Project description

Cloud computing is beginning to accelerate the science process by removing barriers associated with collecting and quality controlling data. Cloud computing also provides the means to improve scientific workflows and promote research sharing through open-source literate programming tools such as notebook (e.g., Jupyter and Rmarkdown). However, this is a seismic shift in how researchers in the Earth science community do their work. Many researchers are resistant to adopting and taking advantage of cloud computing because of the hurdles associated with starting, the lack of domain specific examples, unclear cloud computing costs, and the plethora of cloud computing vendor Application Programming Interfaces (APIs).

There are some existing efforts to create such notebooks to promote the adoption of cloud computing and open-source Artificial Intelligence (AI) tools. However, these efforts are usually side products related to a specific research project and developed by the researchers themselves. The notebook development process typically does not directly engage potential users, which may reduce the value and impact of the final notebooks. In an effort to develop interactive machine learning tutorials supported by ESIP Funding Friday, we found that training materials would be more useful and impactful when potential users were engaged in the development process. Additionally, many existing notebooks do not necessarily follow the best practices in cloud computing and AI applications (e.g., provenance, reproducibility, and content accessibility).

The project proposes creating well-documented notebooks that show how to collect, distribute, process, and analyze geophysical datasets with open-source AI tools. The development process will actively engage potential users to identify learning topics of high demand and seek user feedback along the development process. Additionally, all notebooks will follow and highlight community best practices on cloud computing and AI applications[1]. This project will build a workflow and infrastructure using the open science ecosystem (i.e., Jupyter, Python, R, Google Colaboratory, Binder Project, and GitHub) that is scalable and can enable community contributions with notebook templates, contribution guidelines, and automated evaluation tools.

[1] Rule et al., (2019) Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. PLoS Comput Biol 15(7): e1007007. https://doi.org/10.1371/journal.pcbi.1007007

To demonstrate the diversity of cloud computing resources and public Earth science data, we will develop notebooks that use services and geophysical data from several cloud computing vendor APIs (e.g., Amazon AWS, Google Cloud Storage/Earth Engine) and data sets from various government agencies that have moved portions of their data holdings to the cloud (e.g., NOAA's Big Data Project, NASA's Earthdata Cloud Evolution, USGS's Cloud Hosting Solutions). We will also leverage community-driven tools for open, reproducible, and scalable science, such as the Pangeo software ecosystem, in the notebook development process.

To create notebooks that are relevant to users with different levels of technical background, the project will follow the concept of "learning journey," which is a series of progressive notebooks that are suitable for users with different levels of technical knowledge. The learning journey allows us to separate a complicated learning process into manageable pieces to facilitate more effective learning for potential users. Users can start their own learning journey via different entry points of their choice. The main learning objective of the project team is to identify the best practices and tools to make interactive notebooks accessible to all users by incorporating the Web Content Accessibility Guidelines (WCAG) developed by the World Wide Web Consortium (W3C) (https://www.w3.org/TR/WCAG/).

2.2 Project objectives, significance, and impact

Objectives

In this project, we will develop a series of notebooks focusing on how to use cloud service vendor APIs to collect data, to preprocess data making it both "analysis-ready" and "AI-ready", and tutorials on leveraging various geophysical datasets and probing the data in an AI application context. This series of notebooks will promote the adoption of cloud computing and AI tools in the Earth science community and creating notebooks following community best practices will demonstrate how the cloud assists in research collaboration, dissemination, and reproducibility.

Significance & impact

The impact of this notebook series will promote modern research workflows and assist in enhancing proficiency in AI by demonstrating community-recommended best practices. The increase in quality notebooks will also advance trust and transparency of cloud computing through generating domain specific content, endorsing a notebook structure for research information dissemination, and discussing the costs associated with developing AI applications and analyzing geophysical data in the cloud. This project will also serve as a jumping off point for the Earth science community at large to contribute notebook content following best-practices for literate programming and cloud computing in a consistent approachable manner through the open science workflow and relevant tools.

The infrastructure and materials generated by this project is highly relevant to the technological priorities of federal agencies (e.g., NOAA, NASA, USGS, NSF). Thus, it has the potential to seek additional funding support from various federal agencies, such as, NOAA Center for Artificial Intelligence, NASA ACCESS Program, and NSF Cyberinfrastructure Program, to expand the development and serve the Earth science community.

2.3 Key project steps and timeline

- March 2021 - Community engagement & identifying training topics (1 month)
- April-May 2021 - Data preprocessing (1.5 month)
    - Developing data processing notebooks to make data “AI-ready” and “analysis-ready”
- May-September 2021 - Learning journey development (~2.5 months/journey with 2 journeys)
    - Learning Journey I (3-5 notebooks)
        - Earth science topic: weather/climate forecasting post-processing
        - Dataset(s): NCEP reanalysis dataset on Google Cloud
        - Engaged community: NCAI, NSF AI Institute, UK Met Office, ESIP Machine Learning, ESIP Ag & Climate, ESIP Cloud Computing
    - Learning Journey II (3-5 notebooks)
        - Earth science topic: disaster monitoring and response
        - Dataset(s): GOES-16 ABI dataset on Amazon Web Service
        - Engaged community: NCAI, NSF AI Institute, USGS, ESIP Machine Learning, ESIP Cloud Computing, ESIP Disaster Lifecycle
- October 2021 - Complete and review deliverables (~0.5 months)

2.4 Additional funding currently supporting this work

The work proposed herein will have in-kind benefits from and to the Communities of Practice around AI at NOAA that is being developed by the NOAA Center for AI.

3. **Outreach**

Community engagement & product delivery

The project will engage with diverse institutions and organizations during the tutorial development process with a frequent user feedback mechanism as outlined in Figure 1. The project will seek input from machine learning communities of practice (e.g., NOAA Center for AI, UK Met Office Joint Centre for Environmental Intelligence, NSF AI Institute, and USGS Community for Data Integration) to identify possible topics with high learning demand to guide the learning journey development. Meanwhile, the project team will actively engage with ESIP collaboration areas (e.g., machine learning, cloud computing, agriculture & climate, and disaster lifecycle) to seek feedback on both scientific and technological aspects of the project during notebook development and user feedback process. The developed learning journeys will be shared with the communities of practice and ESIP community to further seek user feedback.

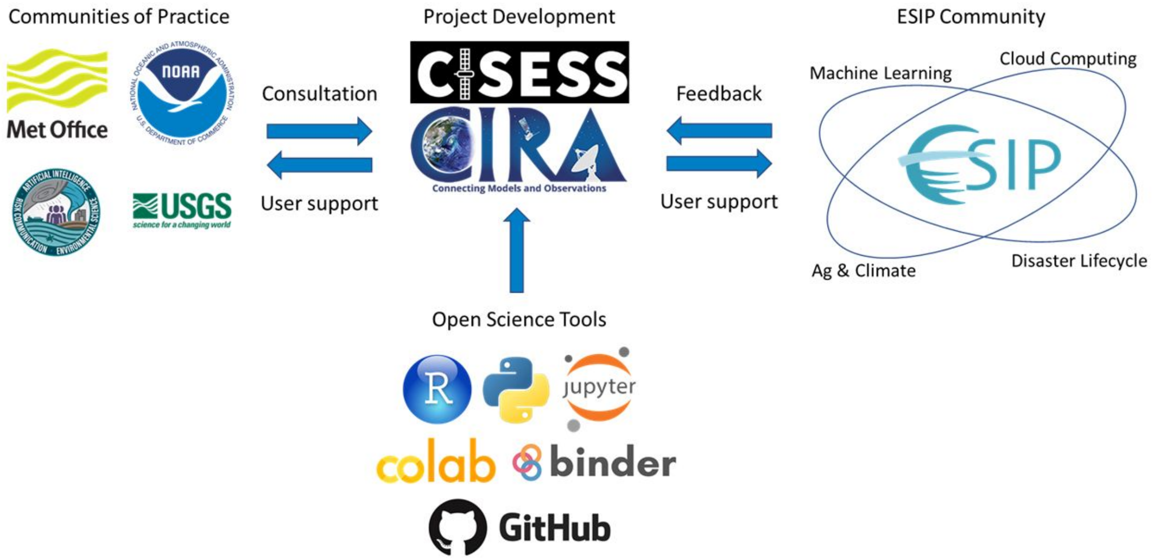

Figure 1. The community engagement framework for the proposed project.

The project will be based on an open science framework as presented in Figure 1. The framework is built on the Google Colaboratory (https://colab.research.google.com/) and the Binder Project (https://mybinder.org/) which are both cloud-based tools for interactive learning and collaboration. The learning journeys will be hosted on a public project GitHub repository. This open science framework will allow everyone to access and share learning journeys generated by the project. Meanwhile, it allows broad community contribution with the potential to become a scalable effort for the Earth science community.

Impact assessment & knowledge sharing

We will assess the impact of the project using the number of users of each tutorial and number of collaborative training events hosted with community partners. The knowledge and experience learned from this project will be shared in an AGU Eos article on how to create cloud-based interactive tutorials to accelerate machine learning training for the Earth science community. Specifically, we will highlight the challenges and experiences of creating interactive tutorials that are accessible to all by following the W3C accessibility guidelines (https://www.w3.org/TR/WCAG/). We will also present the project information and outcome at ESIP meetings, upcoming NOAA AI workshop, AGU, and AMS conferences in 2021.

4. Budget

Community engagement & survey	$1000
Developing data preprocessing pipeline and notebooks	$2000
Developing notebooks for Learning Journey I	$3500
Developing notebooks for Learning Journey II	$3500