PDF Download
3708035.3736089.pdf
14 January 2026
Total Citations: 1
Total Downloads: 803

Published: 20 July 2025

Citation in BibTeX format

PEARC '25: Practice and Experience in Advanced Research Computing
July 20 - 24, 2025
Ohio, Columbus, USA

Conference Sponsors:
SIGHPC

SHORT-PAPER

Toward an Education Hub Linking Research Data and Compute to Learning Workflows in the National Data Platform

MELISSA FLOCA, San Diego Supercomputer Center, San Diego, CA, United States

KATE O'LAUGHLIN, San Diego Supercomputer Center, San Diego, CA, United States

PEDRO RAMONETTI VEGA, University of California, San Diego, San Diego, CA, United States

AMARNATH GUPTA, University of California, San Diego, San Diego, CA, United States

İLKAY ALTINTAŞ, University of California, San Diego, San Diego, CA, United States

MANISH PARASHAR, The University of Utah, Salt Lake City, UT, United States

Open Access Support provided by:

San Diego Supercomputer Center

University of California, San Diego

The University of Utah

PEARC '25: Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration (July 2025)
https://doi.org/10.1145/3708035.3736089
ISBN: 9798400713989

Toward an Education Hub Linking Research Data and Compute to Learning Workflows in the National Data Platform

Melissa Floca[*]
San Diego Supercomputer Center, UC San Diego
La Jolla, CA, USA
mfloca@ucsd.edu

Kate O'Laughlin[*]
San Diego Supercomputer Center, UCSD
La Jolla, CA, USA
kbolaughlin@ucsd.edu

Pedro Ramonetti Vega[*]
University of California San Diego
La Jolla, CA, USA
pramonettivega@ucsd.edu

Amarnath Gupta[*]
University of California San Diego
La Jolla, CA, USA
a1gupta@ucsd.edu

Ilkay Altintas[*]
University of California San Diego
La Jolla, CA, USA
ialtintas@ucsd.edu

Manish Parashar[*]
Scientific Computing and Imaging (SCI) Institute, University of Utah
Salt Lake City, UT, USA
parashar@ieee.org

Abstract

As demand for AI literacy and data science education grows, there is a critical need for infrastructure that bridges the gap between research data, computational resources, and educational experiences. Although national-scale research platforms increasingly provide access to data and compute, integration with educational use cases remains limited. To address this gap, we developed a first-of-its-kind Education Hub within the National Data Platform. This hub enables seamless connections between collaborative research workspaces, classroom environments, and data challenge settings. By leveraging shared infrastructure resources and lowering technical barriers, the Education Hub supports hands-on, data-driven learning at scale. This paper presents the design and implementation of the Education Hub, along with lessons learned from early adopters and case studies that highlight its use in university classrooms and national-scale data challenges. Our findings underscore the value of embedding education directly into the research data ecosystem and point toward future directions for building inclusive, scalable AI and data science education platforms.

Keywords

AI in Science; AI Education; Data Challenges

ACM Reference Format:
Melissa Floca, Kate O'Laughlin, Pedro Ramonetti Vega, Amarnath Gupta, Ilkay Altintas, and Manish Parashar. 2025. Toward an Education Hub Linking Research Data and Compute to Learning Workflows in the National Data Platform. In *Practice and Experience in Advanced Research Computing (PEARC '25), July 20–24, 2025, Columbus, OH, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3708035.3736089

1 Introduction

The rapid expansion of artificial intelligence (AI) and data science has led to growing demand for accessible, hands-on educational opportunities that prepare students and professionals to work with real-world data at scale. However, a persistent gap exists between advanced data and compute resources used in research environments and tools commonly available in educational settings. Although national research platforms offer robust infrastructure for data-driven discovery, their potential to support education remains underdeveloped.

Educators often face barriers when attempting to integrate large-scale data resources or high-performance compute into coursework. These challenges include complex access protocols, limited technical expertise, and a lack of educational tools embedded in the research infrastructure. As a result, students are frequently restricted to simplified or synthetic datasets and isolated learning environments that do not reflect the collaborative, data-intensive workflows of modern research and industry.

To address these challenges, we have developed an Education Hub within the National Data Platform (NDP), a first-of-its-kind initiative that connects classrooms and learners directly to the same data, compute resources and collaborative tools used by researchers. The Education Hub enables instructors to build interactive, data-rich learning experiences and facilitates the design of scalable educational programming such as data challenges, hackathons, and interdisciplinary course materials. It reduces barriers for both instructors and learners by providing pre-configured environments, intuitive interfaces and user experiences, shared access to national-scale resources, and measurement and monitoring tools.

In this paper, we present the motivation, design and implementation of the Education Hub and share insights from its early adoption. We highlight example use cases from university classrooms and national data challenges, showcasing how the platform enables new models for AI and data science education. Finally, we reflect on lessons learned and outline future directions to expand and deepen the role of education within the national data infrastructure.

[*] All authors contributed equally to this research.

PEARC '25, Columbus, OH, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1398-9/25/07
https://doi.org/10.1145/3708035.3736089

2 National Data Platform

The National Data Platform (NDP) [3] is designed to bridge critical gaps in foundational data infrastructure and services by: (1) federating siloed data repositories into a unified platform for discovery, access and use; (2) integrating these resources with advanced computing infrastructure; and (3) providing access and use through standardized processes and customizable services for data ingestion, indexing, curation, and analysis.

Figure 1 shows the current NDP [1] architecture that is designed to generalize, scale and stabilize prior prototype research and education workflows serving multiple communities. At the core of this architecture is a federation capable Hub that not only facilitates discovery of data, compute, research and education resources but also provides extensible platform capabilities toward deployment of a standardized software stack in data and compute endpoints.

This federation architecture built on composability as a principles is couple to a "barrier-removal" approach that combines needs assessment, co-design, and user capacity building to support data providers, scientific and AI researchers, and educators. The result is a suite of accessible services for data discovery, wrangling, and knowledge management that foster collaboration and reduce reliance on ad hoc one-off solutions. The NDP Education Hub, highlighted in Figure 1, was developed through such a co-design effort with educators and data challenge providers to create the necessary building blocks for AI and science education activities. Next we describe these building blocks with motivating examples.

3 NDP Education Hub

The NDP Education Hub empowers educators to seamlessly integrate data and computing resources into learners' workflows and assignments. This is achieved by providing a user-friendly interface for building educational modules that package datasets, models, and source code into ready-to-use environments. These resources are deployed directly within a high-performance computing (HPC) infrastructure, streamlining setup, and eliminating many of the common barriers learners face when accessing data and compute at scale.

The Hub provides portals for: (1) open exploration and discovery of educational resources; (2) formal and open self-guided learner activities; and (3) an educators-only class material discovery, design and deployment capabilities. As shown in Figure 2, these portals are created around a standardized set of features involving education spaces and modules.

The Hub currently provides capabilities for two types of educational spaces:

- **Data Challenges:** NDP data challenges for students and researchers are designed to ensure that we are developing broadly accessible services for education and community building. Participation provides ability to build and access open activities to tackle real-world problems using advanced datasets, computational tools, and AI-driven methods. Data challenge toolkits are developed after each data challenge so that other institutions can easily design their own data challenges to be run through the NDP Education Hub. The toolkits will include the teaching materials developed for each data challenge as well as step-by-step guidance for developing a challenge question that takes advantage of the NDP data ecosystem. Users join data challenges as teams and complete modules as part of the challenge. Access to Data Challenge is public for participation of individual or teams of students with an NDP user account.
- **Classrooms:** Educational spaces for classrooms target creation a library of teachable modules with associated data services. Our main mechanism for developing educational materials is working with educators in their classroom actvities to create teaching- and AI-ready examples of open data and services. Educational spaces allow: (1) instructors to conduct their courses within NDP, incorporating multiple modules, and (2) students to be assigned to groups for collaborative work. Access to Classrooms is given by the instructor.

Both educational spaces depend on **Modules** designed as versatile learning objects that contain data, code, and additional resources that can be deployed and executed in JupyterHub. Modules are intended to provide a hands-on learning experience for learners to jump-start the setup process and get straight into their analysis.

NDP JupyterHub Extension. Learners often face significant hurdles when working with real-world data, including fetching datasets, installing dependencies, managing Git workflows, and collaborating with peers. The NDP JupyterHub Extension streamlines these tasks by enabling one-click access to large datasets, easy installation of *requirements.txt* files, and simplified Git integration. Additionally, the Education Hub supports shared file directories within groups, eliminating the need to transfer large datasets, enabling seamless handoffs between collaborators, and reducing the complexity of Git merge conflicts in group workflows.

High Computing Resource Integration. NDP has partnered with the National Research Platform (NRP) [2] to provide easily accessible and deployable HPC environments for learners at all levels. Educators no longer need to find or provide resources to their learners. The NDP and NRP integration allows educators to build courses with the assurance that learners will have robust and easy-to-use access to HPC resources at their disposal. In addition to accessibility, NRP is free to all students and staff of not-for-profit higher education and community college institutions.

Specialized Views. The Education Hub is specifically organized for the different user types, educators and learners, through the different portals. The Educator Portal is only visible to educators and is their entryway to view all of their created classrooms and data challenges, as well as provide full access to create, update, and manipulate their content. The Learner's Portal provides learners with an easy dashboard of all classroom and data challenges they are associated with. Lastly, the Explore Portal enables all users to discover and join public data challenges.

4 Case Studies and Lessons Learned

To illustrate the impact and practical application of the NDP Education Hub, we now present two case studies that demonstrate its use in both classroom and data challenge settings. These examples highlight how the platform enables real-world, collaborative learning experiences by connecting students directly to research-grade infrastructure. Alongside these case studies, we share early

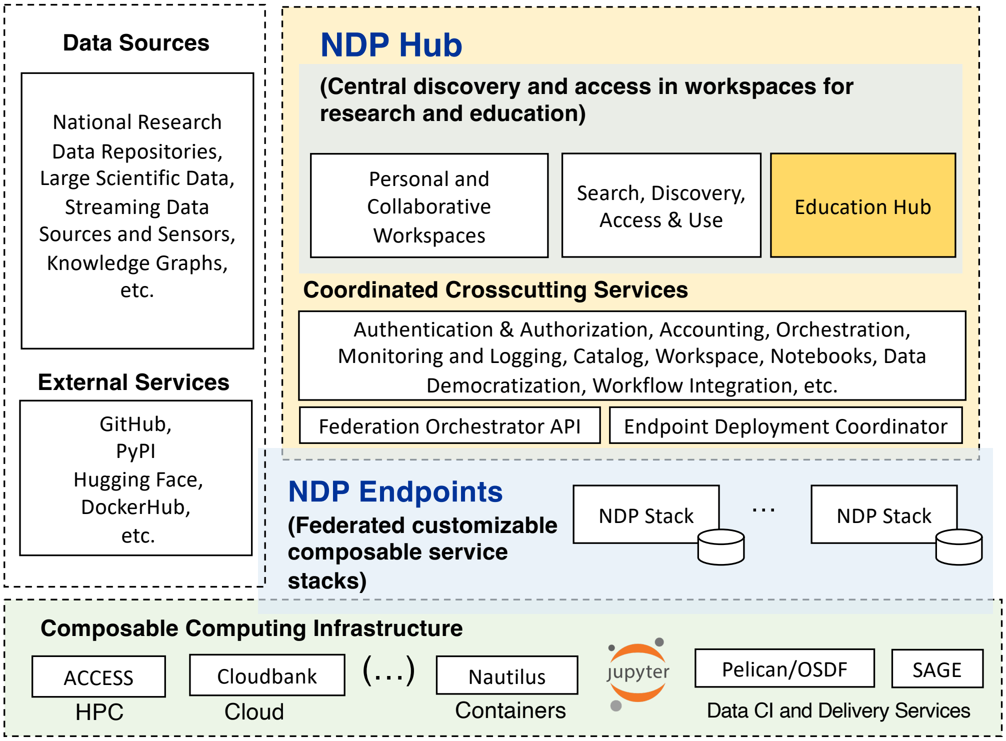

Data Sources
National Research Data Repositories, Large Scientific Data, Streaming Data Sources and Sensors, Knowledge Graphs, etc.
External Services
GitHub, PyPI Hugging Face, DockerHub, etc.
NDP Hub
(Central discovery and access in workspaces for research and education)
Personal and Collaborative Workspaces
Search, Discovery, Access & Use
Education Hub
Coordinated Crosscutting Services
Authentication & Authorization, Accounting, Orchestration, Monitoring and Logging, Catalog, Workspace, Notebooks, Data Democratization, Workflow Integration, etc.
Federation Orchestrator API
Endpoint Deployment Coordinator
NDP Endpoints
(Federated customizable composable service stacks)
NDP Stack
...
NDP Stack
Composable Computing Infrastructure
ACCESS
HPC
Cloudbank
Cloud
(...)
Nautilus
Containers
jupyter
Pelican/OSDF
SAGE
Data CI and Delivery Services

Figure 1: NDP's federated architecture enables bridges existing composable computational and data resources to research and education workflows through both the NDP Hub and customizable service endpoint deployments.

lessons learned from educators and learners that inform ongoing improvements and future directions for the Education Hub.

4.1 Information Integration Classroom Case Study

The NDP platform was put to test in an educational context at a graduate class for Data Science and Engineering students at UC San Diego. This student cohort primarily consists of individuals who are already working professionals, and participate in this graduate program part-time. The class on Information Integration has a class project in which different student groups were tasked to combine information from multiple data systems and integrate the results into a knowledge graph. The different data resources used in the project were registered in the NDP catalog. Each group used the NDP workspace to access these sources and run their integration code using Jupyter Notebooks in the workspace. The workspace also created a separate Neo4J instance for each project group to store their final knowledge graph for subsequent reuse. For several projects, third-party libraries like Memgraph were installed in the workspace and the high-performance computing infrastructure provided by the National Resource Project (NRP) was invoked for data-intensive components of the code, such as semantic matching. Intermediate results of their multi-stage computation were saved in the local temporary storage associated with their workspace. Some groups also used additional data sources stored on their local systems in conjunction with data available via the NDP catalog.

Lessons Learned. Several insightful usage patterns emerged from this classroom experience. First, the primary motivation for using NDP in this class was that student groups find it very difficult to share large and complex data and code from their individual machines; so a platform like NDP serves as a collaborative work-sharing platform for them. In this regard, the NDP platform was a success. However, we also discovered that the platform needed several improvements in other areas. Most students often run long, sometimes overnight, jobs on the NRP platform, and the execution times for these jobs are very hard to estimate – hence the preemptive task scheduling rules need to be configured for this kind of workload. Second, the memory available to the NDP workspace is sometimes insufficient for data-intensive computations. Research is underway to predict resource needs for a specific workflow based on job parameters so that the system can inform the user about potential resource constraints before their jobs are executed. Finally, there is a need for additional documentation including "How to" manuals to accommodate students who are not fully familiar with such platforms because their job environment does not provide any exposure to distributed data/computation infrastructures.

4.2 Fire-Ready Forest Data Challenge

The Fire-Ready Forests data challenge brought together students to advance proactive solutions to end destructive wildfires. Prescribed fires can effectively reduce fuels to restore fire-dependent

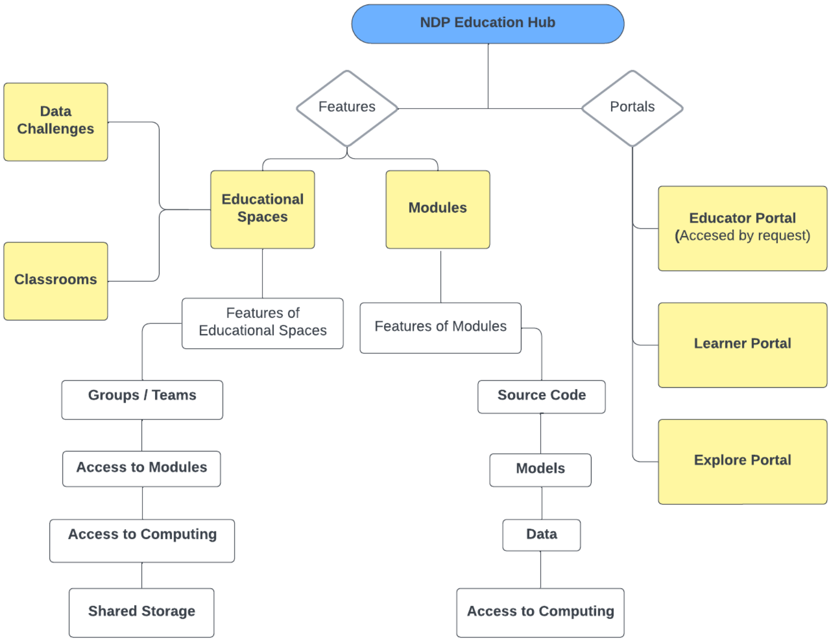
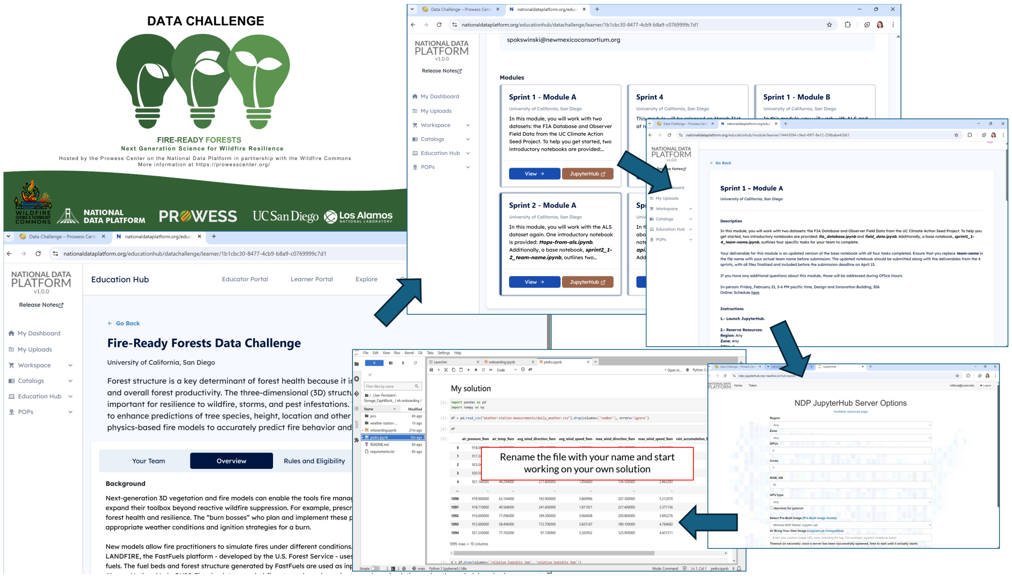


Figure 2: The NDP Education Hub diagram outlines the architecture of our Data Challenge and Classroom features, functionality, and the different portal views to interact and explore the Education Hub.



Hosted by the Prowess Center on the National Data Platform in partnership with the Wildfire Commons
More information at https://prowesscenter.org/

nationaldataplatform.org/educationhub/datachallenge/learner/1b1cbc30-8477-4cb9-b8a9-c0769999c7d1

Fire-Ready Forests Data Challenge

University of California, San Diego

Forest structure is a key determinant of forest health because it i
and overall forest productivity. The three-dimensional (3D) structu
important for resilience to wildfire, storms, and pest infestations.
to enhance predictions of tree species, height, location and other
physics-based fire models to accurately predict fire behavior and

Background

Next-generation 3D vegetation and fire models can enable the tools fire manag
expand their toolbox beyond reactive wildfire suppression. For example, prescr
forest health and resilience. The "burn bosses" who plan and implement these p
appropriate weather conditions and ignition strategies for a burn.

New models allow fire practitioners to simulate fires under different conditions.
LANDFIRE, the FastFuels platform - developed by the U.S. Forest Service - uses
fuels. The fuel beds and forest structure generated by FastFuels are used as inp

spokswinski@newmexicoconsortium.org

Sprint 1 - Module A
University of California, San Diego
In this module, you will work with two datasets: the FIA Database and Observer Field Data from the UC Climate Action Seed Project. To help you get started, two introductory notebooks are provided:...

Sprint 4
University of California, San Diego

Sprint 1 - Module B
University of California, San Diego

Sprint 2 - Module A
University of California, San Diego
In this module, you will work with the ALS dataset again. One introductory notebook is provided: *ttops-from-als.ipynb*. Additionally, a base notebook, *sprint2_1-2_team-name.ipynb*, outlines two...

nationaldataplatform.org/educationhub/module/learner/14443094-c9ed-49f7-8e12-258babe42661

Sprint 1 - Module A
University of California, San Diego

Description

In this module, you will work with two datasets: the FIA Database and Observer Field Data from the UC Climate Action Seed Project. To help you get started, two introductory notebooks are provided: *fia_database.ipynb* and *field_data.ipynb*. Additionally, a base notebook, *sprint1_1-4_team-name.ipynb*, outlines four specific tasks for your team to complete.

Your deliverable for this module is an updated version of the base notebook with all four tasks completed. Ensure that you replace *team-name* in the file name with your actual team name before submission. The updated notebook should be submitted along with the deliverables from the 4 sprints, with all files finalized and included before the submission deadline on April 13.

If you have any additional questions about this module, those will be addressed during Office Hours:

In-person: Friday, February 21, 3-6 PM pacific time, Design and Innovation Building, 306
Online: Schedule here

Instructions

1.- Launch JupyterHub.

2.- Reserve Resources:
Region: Any
Zone: Any

My solution

```
[2]: import pandas as pd
     import numpy as np
[3]: df = pd.read_csv("weather-station-measurements/daily_weather.csv").drop(columns=['number'], errors='ignore')
[4]: df
```

Rename the file with your name and start working on your own solution

	air_pressure_9am	air_temp_9am	avg_wind_direction_9am	avg_wind_speed_9am	max_wind_direction_9am	max_wind_speed_9am
4	921.160000	44.294000	277.800000	1.856660	136.500000	2.863283
...	...	...	...	...	...	...
1090	918.900000	63.104000	192.900000	3.869906	207.300000	5.212070
1091	918.710000	49.568000	241.600000	1.811921	227.400000	2.371156
1092	916.600000	71.096000	189.300000	3.064608	200.800000	3.892276
1093	912.600000	58.406000	172.700000	3.825167	189.100000	4.764682
1094	921.530000	77.702000	97.100000	3.265932	125.900000	4.451511

1095 rows × 10 columns

ndp-jupyterhub.nrp-nautilus.io/hub/spawn

NDP JupyterHub Server Options

Region: Any
Zone: Any
GPUs: 0
Cores: 1
RAM, GB: 16
GPU type: Any
□ /dev/shm for pytorch
Select Pre-Built Image (Pre-Built Image Guide): Minimal NDP Starter Jupyter Lab
Or Bring Your Own Image (JupyterLab Compatible): Enter your custom image URL here, including the tag. For example: jupyter/r-notebook:latest
Timeout (in seconds): once a server has been successfully spawned, time to wait until it actually starts

Figure 3: Learners join and participate in completion of challenge modules through standardized workflows in the NDP Education Hub.

ecosystems and mitigate the risk of devastating wildfires. Next-generation 3D vegetation and fire models can enable the new tools fire managers need for real fuel and fire management solutions.

Collecting field data as inputs for ecosystem modeling and monitoring can be very time intensive and cost prohibitive. New approaches that leverage sensed data (e.g., aerial and terrestrial LiDAR) can improve the speed and scalability of gathering crucial information to model wildfires and wildfire resilience. Currently, we lack good methodologies to predict tree species from LiDAR data. During the data challenge, students created on-ramps for LiDAR and vegetation plot data to build 3D models of forests for ingestion into fire models to explore interactions between topography, fuel structure, fuel moisture, and winds, all at high resolution.

Students were provided modules with exploratory notebooks for a diverse set of data sources that are critical to understanding forest structure to predict fire behavior, including Forest Inventory and Analysis data, field-collected data, and aerial and terrestrial LiDAR data. They then learned to generate tree lists from aerial and terrestrial LiDAR data using open source code. With the background and preparation provided through the modules, teams used machine learning approaches to predict the distributions of trees by plant functional type, genus, and species across six different ecosystems. Figure 3 depicts a learner's pathway for completion of one of the modules in the challenge.

Lessons Learned. One of the most important factors for engagement was selecting a meaningful, real-world problem that resonated with students. Framing the challenge around a timely and relevant issue—like wildfire resilience—sparked interest and motivated participation. Close collaboration with domain experts was critical in curating high-quality datasets, models, and exploratory notebooks. Their input ensured that the challenge materials reflected real challenges in the field, which both grounded the work and elevated its value for student learning and for science.

We found that structuring the challenge in progressive "sprints," beginning with simpler tasks and gradually increasing in complexity, helped build student confidence and sustained momentum. This scaffolded approach allowed participants to gain familiarity with tools and data before diving into more complex modeling efforts. Providing a well-tested onboarding module was key. We developed an interactive walkthrough of the platform that students could complete independently before the challenge began. Investing in beta testing for this module paid off. It minimized confusion, reduced the need for one-on-one support, and helped students hit the ground running.

5 Conclusion and Future Work

In conclusion, the NDP Education Hub represents a significant step toward integrating research-grade data and compute infrastructure into educational settings. By lowering technical barriers and enabling collaborative, real-world learning experiences, the Hub supports a growing community of educators and learners. Although the Hub is in its early stages, important lessons were learned toward co-design of new features. Future work will focus on expanding user support through comprehensive tutorials and guidance for computational performance tuning in both educational and challenge workflows. Additionally, efforts will explore assessing learner performance and deploying dedicated education spaces on NDP endpoints to further scale and personalize learning experiences across the platform.

Acknowledgments

NDP is funded by the US National Science Foundation award #2333609. The authors would like to thank the National Data Platform team as well as other team members and collaborators for their collaboration and support. The Fire-Ready Forests Data Challenge was funded by US National Science Foundation award #2341120.

References

[1] 2025. The National Data Platform website. https://nationaldataplatform.org/

[2] 2025. The National Research Platform website. https://nationalresearchplatform.org/

[3] Manish Parashar and Ilkay Altintas. 2023. Toward Democratizing Access to Science Data: Introducing the National Data Platform. In *2023 IEEE 19th International Conference on e-Science (e-Science)*. 1–4. doi:10.1109/e-Science58273.2023.10254930