Building a Gateway Infrastructure for Interactive Cyber Training and Workforce Development *

Lan Zhao†
Research Computing, Purdue University, West Lafayette, IN, USA
lanzhao@purdue.edu

Carol X. Song
Research Computing, Purdue University, West Lafayette, IN, USA
carolxsong@purdue.edu

Larry Biehl
Research Computing, Purdue University, West Lafayette, IN, USA
biehl@purdue.edu

Venkatesh Merwade
Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, USA
vmerwade@purdue.edu

Matthew Huber
Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, West Lafayette, IN, USA,
huberm @purdue.edu

Jing Liu
Agricultural Economics, Purdue University, West Lafayette, IN, USA
liu207@purdue.edu

Uris Baldos
Agricultural Economics, Purdue University, West Lafayette, IN, USA
ubaldos@purdue.edu

Ilya Shunko
HUBzero, University of California San Diego, San Diego, CA, USA
ishunko@sdsc.edu

ABSTRACT

Science gateways provide integrated data and computation support for research, education and online collaboration. To address the emerging shortage of adequately trained students and workforce in science and technology, more projects are utilizing gateways to disseminate training materials and online modules. However, the contents of these online learning modules are mostly limited to static materials, lacking support for dynamic and interactive learning. As part of the NSF funded GeoEDF project, we enhanced the HUBzero platform with novel capabilities to improve the online learning experience. These improvements include developing general purpose online tools for education purpose and implementing integrated data and tool functions in the HUBzero course module so that instructors can create online course outlines that seamlessly combine static teaching materials with dynamic data, tools, and interactive coding environments such as Jupyter Notebook and RStudio.

In this paper we describe in detail the newly augmented HUBzero course environment implemented on MyGeoHub and showcase several projects that have been using these capabilities on MyGeoHub to support a broad spectrum of learning and training activities, including (1) development and delivery of adaptable cyber training modules for teaching undergraduate and graduate students on the FAIR (Findable, Accessible Interoperable and Reusable) science principles and applications in the fields of hydrology and climate sciences; (2) development and delivery of a short course and workshop tutorials on studying global sustainability using a gridded crop modeling system on MyGeoHub; (3) development of next-generation workforce through undergraduate internships in partnership with the Purdue Discovery Park Undergraduate Research Internship (DURI) program; and (4) outreach to middle school students entering 10^{th} and 11^{th} grade in the TOTAL (Turned Onto Technology and Leadership) summer camp, introducing them to geospatial data management and analysis using MyGeoHub cyberinfrastructure.

CCS CONCEPTS

• **Applied computing** → Education; E-learning; • **Social and professional topics** → Professional topics; Computing education; Informal education.

KEYWORDS

Science gateway, Cyberinfrastructure, HUBzero, Cyber training, Workforce development, Interactive online learning environment, MyGeoHub

ACM Reference Format:
Lan Zhao†, Carol X. Song, Larry Biehl, Venkatesh Merwade, Matthew Huber, Jing Liu, Uris Baldos, and Ilya Shunko. 2020. Building a Gateway Infrastructure for Interactive Cyber Training and Workforce Development *. In *Practice and Experience in Advanced Research Computing (PEARC '20), July 26–30, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3311790.3396639

PEARC '20, July 26–30, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6689-2/20/07…$15.00
https://doi.org/10.1145/3311790.3396639

1 INTRODUCTION

With the advent of Big Data challenges in almost every discipline, it is urgent to train the next-generation students and workforces to develop and/or use cyberinfrastructures and software tools/applications to solve complex real-world problems. Science

gateways provide access to advanced data and computation resources and integrate them with an easy-to-use web frontend, making it an ideal platform to support such training and education efforts. But, instead of fully leveraging the online data management and computation capabilities already in science gateways, most existing gateway-based online training systems are functionally restricted and only provide static learning materials for online viewing or download.

For example, the XSEDE user portal[1] provides a large collection of training materials on XSEDE resources and related technologies, most of which consists of presentation PowerPoints, documents, and videos. Another example is SERC[2] (Science Education Resource Center at Carleton College). Founded to improve education in Earth sciences, the SERC web site hosts more than 100 education projects engaging participation in more than 1000 institutions from K-12 to higher education. Each project page consists of units such as teaching notes, case studies, step-by-step instructions, and references in the form of html pages, URL links and screenshots. Students have to download all the example code, set up their own development environment, download and install software needed before being able to follow the instructions posted online. Such static approaches are a great starting point but not sufficient to train the next-generation workforce with the skills it needs.

To address this limitation, we have enhanced the HUBzero gateway platform with novel functions to provide seamless interactive online learning. HUBzero is an open source scalable science gateway platform for scientific collaboration [1]. HUBzero enables creation of dynamic web sites where domain scientists and students can rapidly develop online interactive applications, publish and share with others who can then use these tools and access computational resources on local, regional or national advanced cyberinfrastructures such as XSEDE. HUBzero provides strong support for teaching and education. Its built-in "course" module includes many features for developing and distributing interactive online course materials. HUBzero's support for collaboration, such as group, project, wiki, forum, tagging, review, citation, Q&A, etc., are commonly used in classroom settings. The recent addition of Jupyter Notebook and RStudio/Shiny to HUBzero has made online tool development and publication easier for teachers to develop and share interactive teaching materials; and for students to learn data processing techniques using the programming languages popular in their domains. Funded by the NSF CSSI program, our GeoEDF project [2] further enhanced the HUBzero platform with new features to improve the online learning experience, including general purpose online tools that are widely used in classrooms and integrated data and tool functions in the HUBzero course module so that instructors can create online course outlines that seamlessly combine static traditional teaching materials with dynamic data, tools, and interactive coding environments such as Jupyter Notebook and RStudio. The enhanced course cyberinfrastructure has been deployed on MyGeoHub (mygeohub.org [3]), an instance of HUBzero that is extended with geospatial data management and visualization services.

In the following sections we describe in detail the enhanced HUBzero infrastructure for online learning, and how its deployment on MyGeoHub gateway has been used to support multiple training activities, including courses and workshops on FAIR (Findable, Accessible Interoperable and Reusable) science cyber training and global sustainability. In Section 4 we describe how general-purpose tools were developed and used to support teaching activities around the world. In Section 5 we discuss training activities of undergraduate students using MyGeoHub through internships in which they learned technologies and skills in developing HUBzero tools and online tutorials using Jupyter Notebook and HUBzero course. In Section 6 we describe the use of MyGeoHub platform to engage middle/high school students in STEM education. We conclude the paper in Section 7.

2 ENHANCED HUBZERO INFRASTRUCTURE FOR INTERACTIVE ONLINE LEARNING

Building on the open source HUBzero platform, the MyGeoHub science gateway has served as an online environment for research project hosting, team collaboration, as well as online teaching and workforce development. HUBzero provides out-of-box support for education through its "course" component and supplemental functions such as group, project, wiki and forum that have been used to develop more than 100 courses across a range of technical areas at their respective hubs. Funded by the NSF DIBBS and CSSI programs, MyGeoHub extended and enhanced the HUBzero course environment to support seamless integration of data access and tool launching functions (Figure 1).

Augmenting the traditional static online learning modules, instructors can easily build a course outline that combines multimedia learning materials with hands-on interactive activities such as directly launching an online modeling tool with example input data to demonstrate the usage of the tool as well as how to use it to study a real world problem. Furthermore, by integrating HUBzero course with Jupyter Notebook and RStudio/Shiny environments, instructors can develop and share interactive teaching materials via a combination of explanatory text, skeleton code block, and visual output, all in a single web-based document. Finally, the common underlying geospatial and high-performance computing service infrastructure that has been deployed on MyGeoHub makes it convenient for instructors to develop courses that involve geospatial data and simulations.

These enhancements were driven by user requirements. Instructors and organizers of various teaching and training events have provided input and served as early testers. Continuous feedback and iterative development/deployment cycles have resulted in much improved user experience.

2.1 HUBzero Course and Data Management

HUBzero course comes with many easy-to-use functions. Instructor can create a course landing page with comprehensive information including an overview, a short description, a course image, time and effort, a tab for students to enter reviews, and enrollment options. The built-in authorization function allows a course owner to give access to co-instructors so that multiple people can develop a course at the same time. Inside a course, the main functions can be accessed via tabs for outline, progress tracking, notes, announcements, and discussions. An instructor can create a course outline that consists

[1] https://portal.xsede.org/online-training
[2] https://serc.carleton.edu/

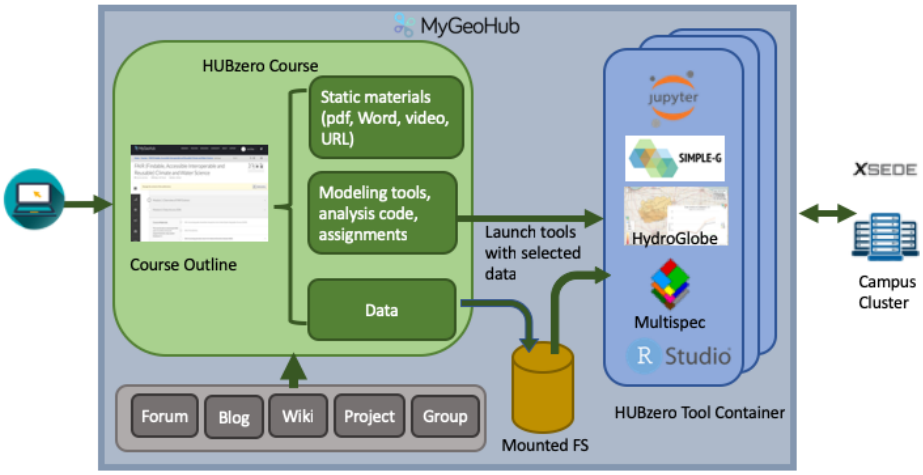

Figure 1: The enhanced HUBzero course infrastructure

of multiple units. Inside each unit, he/she can include lectures, activities such as lab exercise and homework assignment, and an exam. Files of various formats can be uploaded including PDF, MSWord, video, as well as URL links. YouTube and Kaltura videos can be embedded into a course outline. An instructor can also create a wiki page as part of the outline instead of uploading a file. In the case when an instructor needs to share some data files with the class, he/she can upload the data directly from his/her laptop or from one of his/her hub projects which come with its own file storage. The uploaded data can be accessed from the HUBzero tool container where simulation code executes. By integrating course functions with the HUBzero project file space instructors can easily share data with students and launch interactive tools using the shared data.

2.2 Tool as a Service

HUBzero enables domain researchers (e.g., experts in hydrological and climate science) to easily develop and publish their model and data analysis code online by following a step-by-step procedure. The code is run securely in a separate Linux tool container and users interact with the tool's user interface using a browser via VNC (Virtual Network Computing). There is no need for users to download or install any software. In recent years, in addition to the traditional HUBzero tools that run as a desktop application in a tool container, HUBzero added Jupyter Notebook and RStudio/Shiny environments to its tool portfolio, making it even easier for domain scientists to develop and publish tools using the programming languages they are familiar with and the latest software development environment with markdown support to create dynamic documents with embedded code which is very well suited for teaching and learning.

In addition to developing general purpose data access, processing and visualization tools for teaching in multiple disciplines, we worked with the HUBzero team to develop the capability that allows a HUBzero tool to be launched via a URL with invocation parameters. The parameters are saved in an environment file in the workspace. At launch, the tool reads the environment file and handles the input parameters accordingly. For example, the parameter could be an input file that is open and loaded upon launching the tool. This function allows tools to be programmatically launched from other applications on the hub (such as a course or the file browser in a project) or even from another CI system, making tools interoperable across systems.

Enabled by the tool-as-a-service function, when a student clicks on a tool activity in a course, an instance of the hub tool will be launched inside a tool container using the invoke URL. If the instructor attaches certain files with the tool, the file names will be passed as parameters in the URL and will be handled by the tool upon launching. The data selected are stored in a file system that is mounted to the tool container with read access and therefore can be loaded by the tool directly, providing a seamless experience.

Building on the aforementioned comprehensive support for online course development, several projects have created online training modules and courses on MyGeoHub to support their teaching activities. A common usage pattern among them is to form a hub group for the class. The group provides a central dashboard for students to navigate among related modules such as wiki, blog, project, files, and courses. The group owner can self-manage group membership and edit group pages. In the next few sections we highlight several examples on how these functions of HUBzero and enhancements on MyGeoHub have been used by real world projects.

3 CYBER TRAINING AND GLOBAL SUSTAINABILITY COURSES AND WORKSHOPS

3.1 FAIR Cyber Training Course

In the NSF funded Cyber Training for Findable, Accessible, Interoperable, and Reusable (FAIR) Science project (fairhub.org), researchers are addressing the challenges associated with the growing

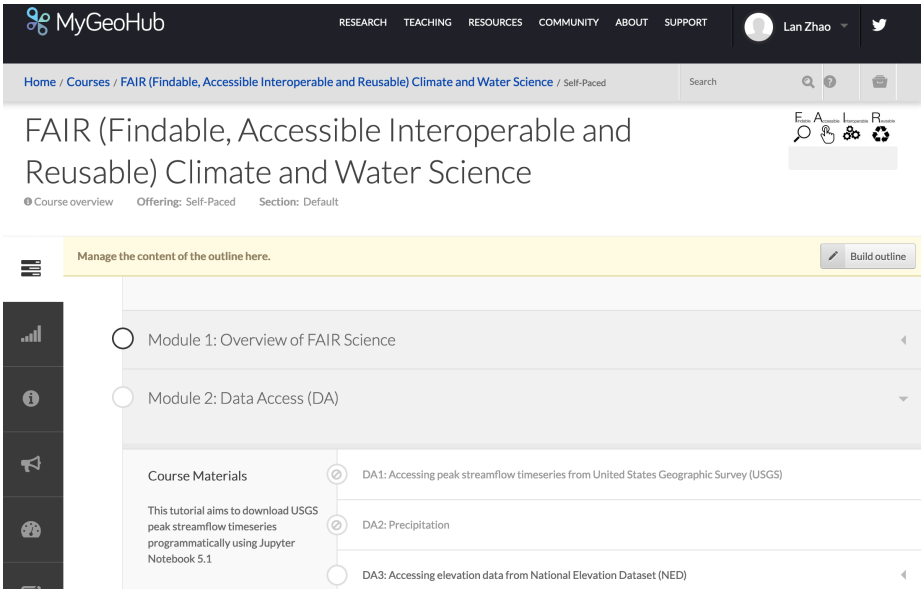

Figure 2: A screenshot of FAIR Climate and Water Science online course on MyGeoHub.

population, food and water security, natural disasters, and changing climate by empowering the next-generation workforce with the necessary cyber skills in FAIR science practice to work with big data analytics and simulations. Its approach is to develop adaptable cyber training learning modules for teaching undergraduate or graduate level students on the FAIR science principles and applications in the fields of hydrology and climate sciences. The learning modules are developed in a flexible way so that they can be easily adapted to other disciplines in the future. These modules are designed to be delivered in multiple formats such as semester course, summer bootcamp, or workshop tutorial at Purdue, University of New Hampshire, and University of Alabama.

In the Fall 2019 semester, the investigators of the cyber training project developed and taught a suite of FAIR-compliant data modules in a graduate course EAPS 59100 (FAIR Data Practice Climate Science) at Purdue University where the students learned the basics of data access, processing, modeling, visualization, and publishing following the FAIR principles and, as a homework, applied their knowledge in their own research projects. The instructors and students used MyGeoHub platform extensively during the class. A FAIR cyber training group was created for the class with a HUBzero wiki being used for discussion and assignments and project file space being used for data sharing. The students also used the Jupyter Notebook environment to develop their data access/processing code directly online and published some of their data and models with Digital Object Identifiers (DOIs) on MyGeoHub.

Following the success of EAPS 59100, two courses are being created using the content taught in the Fall 2019 semester targeting graduate and undergraduate students respectively. Figure 2 shows a screenshot of the FAIR Climate and Water Science course[3] being developed on MyGeoHub. It consists of lecture notes as well as lab activities using Jupyter Notebook. The instructor provides example geospatial data files and skeleton Jupyter Notebooks that can be directly launched and opened in a hub tool container by clicking on the tool activity links in the course outline. Students can follow the instructions to enter their code segments that perform various data accessing and processing tasks and then run it to see the result. Students do not need to download any data or install any tools or libraries by themselves which could get complex and take a lot of time. Instead they can focus on learning the data analysis techniques. This online course is scheduled to be completed by the end of April and used in a workshop in May to teach instructors from institutions outside of Purdue University to adapt this online course for their own science domain and integrate it into their own curriculum.

3.2 Short Course and Workshops on Global Sustainability

To address the urgent need of meeting the Global Sustainable Development Goals on a changing planet with limited land and water resources, the GLASS[4] (Global to Local Analysis of Systems Sustainability) project draws heavily on the data and computation capabilities of MyGeoHub in research, teaching, and collaboration. An interdisciplinary team of GLASS researchers created an online *Short Course in Multi-scale Analysis of Sustainability*[5] and used it to teach an intensive short course that combines online elements taken from an existing agriculture economics course (AGEC 528) with one week of face-to-face interactions on the Purdue campus during the week of September 16, 2019. Approximately 30 attendees came from US and international universities (some in an NSF-funded INFEWS project) and used MyGeoHub throughout the week. In addition to learning new gridded modeling methods, they also gained

[3]https://mygeohub.org/courses/fair_climate

[4]https://www.purdue.edu/project-glass

[5]https://mygeohub.org/courses/sustainability_shortcourse

Figure 3: Researchers learned about global gridded modeling enabled by the MyGeoHub platform at a sustainability training workshop.

hands-on experience in studying real world sustainability challenges and potential policy impacts using an online modeling tool called SIMPLE-G-US[6] available on MyGeoHub and linked through the short course.

Similar training events were also held in the Long Run Sustainability of US Agriculture Conference in September 2018, the Workshop on Open-Source Analysis of SDGs at the Food-Water-Energy Nexus Using Global, Gridded Modeling in November 2018, the Workshop on Open-Source Analysis of SDGs at the Food-Water-Energy Nexus Using Global (Figure 3), Gridded Modeling at Tecnológico de Monterrey in Mexico in June 2019, and the 2019 Sustainability and Development Conference in October 2019, all of which were conducted using the SIMPLE-G-US economic modeling tool and the MyGeoHub platform. These training and outreach events helped broaden the adoption of the SIMPLE-G gridded modeling framework in the community and disseminate the GLASS team research results effectively to policy makers, farmers, and other stake holders.

4 USE OF GENERAL-PURPOSE TOOLS IN TEACHING

Several general-purpose tools, such as MultiSpec Online [5], AgMIP Data Aggregator [6] and HydroGlobe [7], have been developed and published on MyGeoHub using the HUBzero tool framework. These tools support common tasks related to data access, processing, and visualization for geospatial data and domain science modeling outputs. They have been used in classes from institutions around the U.S. in the past few years.

MultiSpec Online (Figure 4) is one of the most widely used tools in teaching and learning on MyGeoHub. MultiSpec is a freeware image data analysis application developed for interactively analyzing Earth observational multispectral and hyperspectral image data from airborne and spaceborne systems, as well as a number of other types of multispectral image data, such as medical images on MacOS and Windows platforms. There are currently in excess of several thousand known, registered users, from a wide range of sectors (research, higher ed, K-12, government, etc.), who download and install the software package on their own computers. MultiSpec has also been integrated into MyGeoHub (named MultiSpec Online) using HUBzero's tool infrastructure to provide users with easier access and a more seamless experience in a web browser. MultiSpec is programmed in C/C++; the wxWidgets library is used for the GUI interface to run on Linux on the HUBzero platform. The software uses several other libraries such as gdal, hdf4, hdf5, netcdf and OpenJPEG to handle the many different file formats for remote sensing image data. The software continues to attract users, many of whom use the software in their classrooms for education and training. Examples have been a NASA sponsored internship for 5 high school students at the University of Texas at Austin and in geospatial training sessions for 25-30 middle school students at Purdue University. Other institutions that have used MultiSpec Online in classrooms include South Dakota State University, Royal Melbourne Institute of Technology in Australia, University of Wisconsin Stout, Kennesaw State University in Georgia, and Nova Southeastern University, to name a few. We have observed a rapid uptake of MultiSpec Online, tripling in the past three years with a total of more than 1000 users to date who have used it in classrooms and research.

5 UNDERGRADUATE INTERN PROGRAM

More than nine undergraduate students have been trained in various projects using the MyGeoHub platform via the Discovery Park

[6]https://mygeohub.org/tools/simpleus

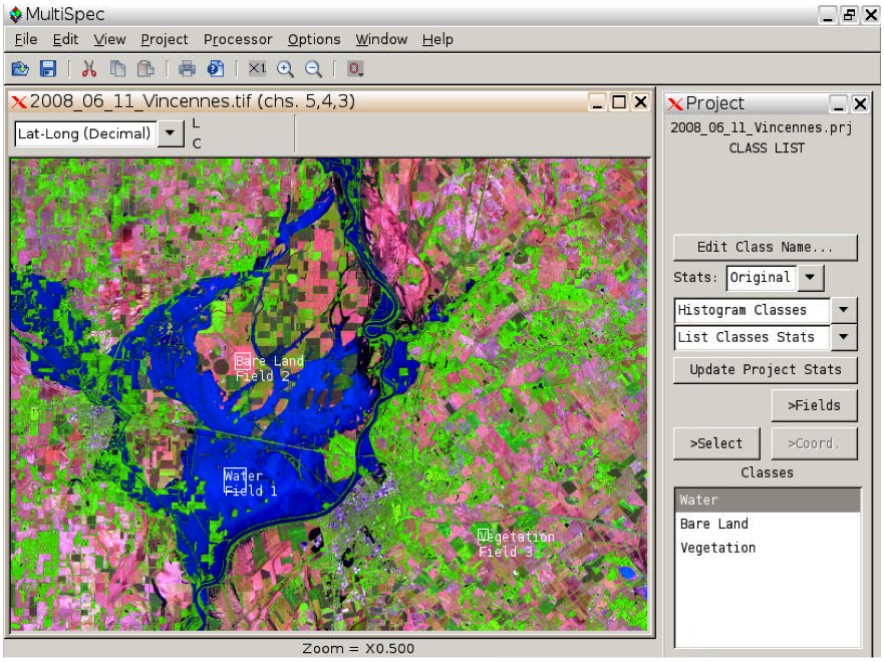

Figure 4: The MultiSpec Online tool for image processing and visualization illustrating the Landsat 5 image of southwestern Indiana and southeastern Illinois that was flooded during June 2008.

Undergraduate Research Internship program at Purdue University. The students worked with researchers in multiple disciplines to utilize scientific data in domain science computational models and data analytics, including urban resilience, sustainable agriculture and water resources. They prototyped and developed software modules and applications on MyGeoHub and contributed to multiple tools and software components currently in production use including MultiSpec Online, the AgMIP data aggregator, SWATShare [8], HydroGlobe, SIMPLE-G-US, as well as a number of geospatial data services deployed on MyGeoHub.

As an example, a freshman student from the computer science department, developed a self-paced online short course titled "Short Course in Geospatial Data Extraction and Visualization" in summer 2019. This course teaches learners from domain scientists to high school students how to develop Jupyter Notebook code to display and explore geospatial data in different formats, including shapefile, geoJSON, and geotiff, and in different sizes (Figure 5). The course consists of four modules and each module consists of a lecture unit and an activity unit. The lecture unit teaches some background information on the various geospatial data formats and how to write Python code to process the data, while the activity unit includes a Jupyter Notebook tutorial and some sample geospatial data students can work on. Lu's work also serves as a use case that exercises the new features enhanced by the GeoDEF project such as the integration of Jupyter Notebook and Project file space with HUBzero course. This course[7] has been used to help developers from another NSF funded project to implement geospatial data visualization functions in their online system.

[7]https://mygeohub.org/courses/geo_viewing

Another junior computer science student is currently working on enhancing the job management and output visualization interfaces for the SIMPLE-G-US tool. She is implementing database-backed management of private and public jobs, retrieval of corresponding job results, as well as extraction of raster data based on user's selection on a map for visualization. Through the intern program, she is able to learn the full development cycle of an NSF-funded project as well as to gain software development skills on scientific data processing and visualization on MyGeoHub.

In addition to learning software engineering and cyberinfrastructure skills, these computer science students have gained experience working in an interdisciplinary team, exposed to requirements from domain science applications that they are not typically in contact with through their normal course of undergraduate studies, an important step in readying them for the future workforce.

6 TOTAL CAMP

MyGeoHub and its general-purpose online tools have also proved to be an excellent platform to engage k-12 students in work with data in ways that will better prepare them for learning and research in college and future jobs.

In the summers of 2016 and 2017 30 to 36 8^{th}- 10^{th} graders of diverse backgrounds from around the United States learned some basic knowledge and techniques on remote sensing and geospatial data analysis at Turned Onto Technology & Leadership (TOTAL) camps at Purdue University (Figure 6). The students gained hands-on experience in using MyGeoHub and the MultiSpec Online tool to investigate some real-world problems, such as to determine the area of a flood in southern Illinois and Indiana which occurred in

Figure 5: A Purdue undergraduate student developed an online course to teach new users how to develop code to visualize geospatial files.

Figure 6: Students in TOTAL Camp learning about geospatial data and analysis using MyGeoHub.

June 2008 using Landsat 5 multispectral satellite data. The students also gained experience in finding locations within Indiana with the highest reported rainfall events using the tools available on MyGeoHub. Valuable feedback was collected at the end of the camp to improve the tools and MyGeoHub.

7 CONCLUSION

We have described our work in enhancing the HUBzero framework to support active online learning. The integration of data, modeling, and interactive software development environments such as Jupyter Notebook and RStudio into the HUBzero Course module, in combination with a comprehensive set of social and collaboration functions such as group, project, wiki, forum, blog, and file sharing, makes HUBzero a unique platform for online education. We also described how the developed infrastructure is being used in multiple projects on MyGeoHub for a variety of training and education activities, reaching out to a wide spectrum of audiences including K-12 students, undergraduate and graduate students, researchers, and policy makers. Through this work, we have learned that it is important to improve the user experience for both end users/learners and instructors in order to gain adoption; a seamless environment linking learning materials, data and tools significantly improves the teaching and learning experience; and an open and self-manage platform such as MyGeoHub has great potential in reaching a broader audience.

ACKNOWLEDGMENTS

This project is funded by the NSF awards no. 1261727, 1829764 and 1835822.

REFERENCES

[1] M. McLennan, R. Kennell, HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering, Computing in Science and Engineering 12 (2) (2010) 48-52.

[2] Kalyanam, R., Zhao, L. and Song, X.C. **GeoEDF – An Extensible Geospatial Data Framework for FAIR Science**.

[3] R. Kalyanam, L. Zhao, C. Song, L. Biehl, D. Kearney, I.L. Kim, J. Shin, N. Villoria, V. Merwade, **MyGeoHub - A sustainable and evolving geospatial science gateway**, Future Generation Computer Systems (2018), *doi.org/10.1016/j.future.2018.02.005*.

[4] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[5] Larry Biehl, *et al.* **MultiSpec: A Desktop and Online Geospatial Image Data Processing Tool**, oral presentation, AGU Fall Meeting, New Orleans, LA, Dec 2017.

[6] Villoria, NB, Elliott, J, Müller, C, Shin, J, Zhao, L, and Song, C., 2016, January. Rapid aggregation of global gridded crop model outputs to facilitate cross-disciplinary analysis of climate change impacts in agriculture. Environmental Modelling & Software. 75, pp.193-201. doi:10.1016/j.envsoft.2015.10.016.

[7] Adnan Rajib, Venkatesh Merwade, Lan Zhao, Jaewoo Shin, Jack Smith and Carol Son. **HydroGlobe – A cyber-enabled platform for auto-extraction and processing of earth observations for hydrologic analysis**, *2017 CUAHSI Conference on Hydroinformatics*, July 25 – 27, 2017, Tuscaloosa, Alabama.

[8] Rajib, M.A., Merwade, V., Kim, I.L., Zhao, L., Song, C. and Zhe, S., 2016, January. SWATShare–A web platform for collaborative research and education through online sharing, simulation and visualization of SWAT models. *Environmental Modelling & Software*, *75*, pp. 498-512. *doi:10.1016/j.envsoft.2015.10.032*.