https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

WaterSoftHack Cybertraining: Reproducible Data Science, Machine Learning, and Cloud and Edge Computing Training for Collaborative Water Science Research

Krishna Panthi[1], Vidya Samadi[2,3*], Nima Zafarmomen[2], Mostafa Saberian[4], Adarsha Neupane[2], Carlos Erazo Ramirez[5], Bijaya Adhikari[6], Anthony Castronova[7], Ibrahim Demir[5,8]

1. School of Computing, Clemson University, Clemson, SC, USA
2. Department of Agricultural Sciences, Clemson University, Clemson, SC, USA
3. Artificial Intelligence Research Institute for Science and Engineering (AIRISE), School of Computing, Clemson University, SC, USA
4. The Glenn Department of Civil Engineering, Clemson University, Clemson, SC, USA
5. ByWater Institute, Tulane University, New Orleans, USA
6. Department of Computer Science, University of Iowa, IA, USA
7. Consortium of Universities for the Advancement of Hydrologic Science Inc, MA, USA
8. River-Coastal Science and Engineering Department, Tulane University, New Orleans, USA

*corresponding author: samadi@clemson.edu

Highlights

1. Hands-on cybertraining in data science, machine learning for water science.
2. Open-source WaterSoft Python package enables reproducible workflow development.
3. Participants show skill gains, collaboration, and applied hydrological research impact.

Abstract

The growing complexity and volume of data in water science demand advanced computational skills among researchers, yet significant barriers limit rapid skill acquisition. We present WaterSoftHack, a two-week cybertraining program designed to equip students, early-career professionals, and researchers with reproducible data science, machine learning, and cloud/edge computing skills for water science applications. The program combines open-access training resources, including the WaterSoft Python package, with cohort-based capstone projects. A rigorous selection process of interested candidates ensures diverse participation across academic levels, institutions, and backgrounds. The training integrates immersive, hands-on instruction with formal science communication, emphasizing reproducibility, scalability, and teamwork. Drawing on surveys and qualitative interviews from the first two years, we demonstrate notable skill advancement, collaborative synergy, and career advancement outcomes. WaterSoftHack highlights the importance of project-based, integrated cybertraining in building computational capacity and preparing a diverse, capable workforce for the data-driven future of water science and engineering.

Keywords: WaterSoftHack; Cybertraining; Project-based Learning; Workforce Development.

Short Summary

WaterSoftHack is a cybertraining program that bridges water science and modern computational methods through hands-on training in data analytics, machine learning, and cloud and edge computing. Using open-source tools and project-based learning, the program builds reproducible workflows, strengthens interdisciplinary collaboration, and enhances participants' research capacity and career readiness.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

1. Introduction

Traditional hydrological modeling has long relied on conceptual and physically based models that represent dominant watershed processes, such as rainfall–runoff dynamics and storage mechanisms (Seibert and Bergström, 2022).. Early pioneers such as Bjerknes (1904) envisioned weather prediction as solving differential equations for fluid dynamics and thermodynamics, while Richardson (1922) articulated a numerical step-by-step procedure for weather forecasting, which remained impractical for decades. It was Charney et al. (1950) who proved the feasibility of digital numerical weather prediction (NWP) experiments via experiments on the ENIAC machine. In parallel, hydrological process modeling advanced with the Saint-Venant equations (Saint-Venant, 1871) for open channel flow and later the kinematic wave theory of Lighthill & Whitham (1955), which provided efficient formulations for flood and storm runoff routing. Empirical rainfall-runoff models like unit hydrographs (Sherman, 1932) and more general instantaneous unit hydrographs (Nash, 1957) further shaped the numerical modeling of early operational hydrology. By the late twentieth century, integrated physically based frameworks such as Modelling Integrated Kinetic Environment – Système Hydrologique Européen (SHE/MIKE-SHE; Abbott, et al., 1986), land-surface hydrology models like Variable Infiltration Capacity (VIC; Liang et al., 1994), and terrain-driven conceptualizations like TOPography-based hydrological MODEL (TOPMODEL; Beven & Kirkby, 1979) broadened the modeling toolkit from catchment to continental scales.

Alongside modeling advances, the acquisition of hydrologically relevant data transformed radically. Gauges provided the first temporally resolved precipitation and discharge records, while the radar networks provide near-real-time rainfall estimates at kilometer and minute scales (Seo et al., 2019). Satellite missions such as Tropical Rainfall Measuring Mission (TRMM; Kummerow et al., 1998) and Global Precipitation Measurement (GPM;Hou et al., 2014) offer near-global precipitation at a few hourly resolutions. In addition, Soil Moisture Active Passive (SMAP; Entekhabi et al., 2010) retrieves global soil moisture and Gravity Recovery and Climate Experiment (GRACE; Tapley et al., 2004) reveals total water storage changes as an effect to earth's gravity. Reanalysis products such as the National Centers for Environmental Prediction (NCEP)/ National Center for Atmospheric Research (NCAR) (Kalnay et al., 2018) and Fifth Generation of ECMWF Atmospheric Reanalysis of the Global Climate (ERA5;Hersbach et al., 2020) provide consistent, gridded meteorological forcing fields at hourly resolution. The data landscape available to hydrologists today is richer than ever, spanning ground-based, airborne, and spaceborne observations (e.g., Demir et al., 2015).

However, this abundance of data introduces a new challenge. Classical hydrology curricula remain primarily focused on physical process theory and mathematical formulations, while training in the computational cyberinfrastructure required exploiting modern computational skills which are limited in hydrology. Students in hydrology learn the governing equations of infiltration, evapotranspiration, or groundwater flow, but rarely acquire skills in managing terabyte-scale datasets, developing reproducible workflows, or deploying scalable models on cloud or edge platforms. This mismatch is becoming increasingly critical as hydrology transitions into data- and computation-rich science.

The gap can be described as threefold. First, although hydrological data from satellites, radars, and reanalysis are openly available, researchers often lack the appropriate skills to acquire, preprocess, and integrate these heterogeneous datasets. Second, while data science and machine learning have exploded in Earth sciences (Krajewski et al., 2021), they are often encountered by hydrology students only superficially,

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

without guidance on their strengths, limitations, and integration with physical process understanding. Third, modern hydrological prediction and water resource management increasingly require high-performance or cloud/edge-based computational infrastructure; for example, collecting sensor monitoring data for running distributed hydrology models such as Weather Research and Forecasting – Hydrological modeling system (WRF-Hydro; Gochis et al., 2015) or producing ensemble forecasts (e.g., Demargne et al., 2014) for understanding flood risk. However, most hydrology curricula lack advanced computing and do not train students to use these models.

This gap has tangible consequences. Operational agencies such as the U.S. National Weather Service rely on ensemble forecasting systems running on supercomputers and assimilating multiple streams of satellite and in situ data (Brown et al., 2014). Research repositories such as NASA's Earth data distribute petabyte-scale archives that are impractical to download, requiring direct cloud/edge-based analysis instead. Without training in data acquisition frameworks, data science methods, and modern computing platforms, new generation of hydrologists risk being unprepared to fully exploit available resources or to contribute effectively to operational and policy-relevant decision-making.

To address this gap, targeted training initiatives are needed to explicitly integrate computational methods into water science and engineering education. Such programs should cover the modern computational stack including (1) data acquisition and management for hydrological datasets, (2) data science principles including statistical learning and visualization, (3) machine learning algorithms for both image and time series pattern discovery and hybrid modeling, and (4) deployment on high-performance, cloud/edge infrastructures. Most importantly, these approaches should be taught in a way that connects directly to hydrological applications and emphasizes reproducibility, scalability, and collaborative practice.

Training workshops and hackathons have emerged as a way of bridging this gap. Hackathons provide a project-based learning opportunity. They present challenges to participants that mirror real life scenarios, often requiring skills not taught in academia (Blumenfeld et al., 1991). Hackathons have recently emerged as innovative platforms for skill development, fostering collaboration, problem-solving, and rapid prototyping (Haw and Crawford, 2025). Research by Briscoe (2014) and Lara & Lockwood (2016) show that hackathons are effective in promoting interdisciplinary teamwork and accelerated learning. Similarly, training workshops have been shown to be pivotal in fostering continuous learning and skill development in hydrology, as they provide immersive, hands-on experiences with emerging methods, tools, and interdisciplinary practices that help professionals and students stay current in this rapidly evolving field (Huppenkothen et al., 2017). Indeed, hackathons with dedicated data-focused tracks provide participants with valuable experience in handling real-world datasets and addressing common issues such as data reliability, error assessment, and reproducibility (Anslow et al., 2016). Henceforth, participants are challenged to apply newly acquired skills to water-related projects, ensuring that the training directly translates to practical applications.

The WaterSoftHack program is an innovative hackathon-style cybertraining initiative that seamlessly integrates data science, machine learning, and cyberinfrastructure principles. By combining hands-on exercises with project-based learning, the program equips researchers with practical skills in data-driven analytics and a deep understanding of the supporting cyberinfrastructure necessary for modern water research. This study presents the outcomes of WaterSoftHack cybertraining program, combining one week

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

of theoretical and hands-on workshops with one week of capstone projects, during which participants tackled real hydrological problems using state-of-the-art data and computational tools.

Participants included undergraduate and graduate students and early-career researchers with diverse backgrounds in hydrology, many of whom had limited prior exposure to computational methods. By documenting the program's structure, delivery, and evaluation, we aim to contribute to the broader discussion on training the next generation of water research scientists. Specifically, we: (a) identify key areas of computational knowledge currently underrepresented in hydrology education; (b) demonstrate how intensive short-term training can enhance participants' competencies in data acquisition, data science, and cloud/edge-based computation; and (c) highlight the challenges encountered and lessons learned in designing interdisciplinary training that bridges hydrological science with modern computational practice.

This paper is organized as follows: Section 2 describes the WaterSoftHack program, including participant recruitment and curriculum development. Section 3 outlines the project design and evaluation methodology. Section 4 presents the results and discussion, and Section 5 concludes the paper with key findings and insights.

2. WaterSoftHack Program Description

WaterSoftHack was conceived as a cybertraining hackathon aimed at accelerating the adoption of data analytics, machine learning, and cloud/edge-based workflows in water science. Its underlying theory of change posits that structured, mentored, hands-on exposure to reproducible computational methods would (i) lower barriers to adoption, (ii) catalyze open-source contributions, and (iii) foster cross-disciplinary collaboration among water science researchers. The program achieves this by providing accessible instruction (educational empowerment), enabling open-source workflow development (research innovation), cultivating a workforce proficient in advanced cyberinfrastructure (collaborative synergy), and promoting broad community adoption of these approaches (community engagement).

2.1. Program Organization and Recruitment

In WaterSoftHack cybertraining, instruction was delivered by faculty and researchers with expertise in hydroinformatics, data science, and computer science, while graduate students mentored and facilitated small-group activities and provided technical support. Eligible participants included undergraduates, graduate students, postdoctoral researchers, and early-career faculty, selected from a large and diverse applicant pool. Recruitment opened in February via social media (e.g., LinkedIn) and professional channels such as the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) website and newsletters. Two months window was provided for applicants to fill in the application form. After the application deadline, applicants were collectively evaluated based on technical experience, training needs, alignment with program goals, and overall potential. The program selected a coherent of WaterSoftHack Fellows each year. We intentionally sought diversity across different criteria when selecting the Fellows. The criteria included institution types (very high research activity (R1), high research activity (R2) , Primarily Undergraduate Institution (PUI), Hispanic-Serving Institution (HIS), and Historically Black College or University (HBCU)), academic roles, gender, disability status, and disciplinary backgrounds. In Year 1, 15 Fellows were selected to participate and compete in the hackathon. This number was increased to 18 in Year 2 (2025), with several Year-1 Fellows continuing. The program attracted a broad set of audiences, with over 150 general participants in 2024 and over 200 in 2025. The selected

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

fellows conducted capstone projects and competed for a prize, forming the core of the WaterSoftHack Fellows program.

2.2. Program Structure and Curriculum Development

The training program was designed as an intensive, two-week online event hosted in late July and early August. It was delivered synchronously via Zoom, with a daily schedule from 11:00 to 15:00 U.S. Eastern Time, incorporating a one-hour lunch break. This format provided approximately 30 contact hours, emphasized hands-on practical training (60%) combined with theoretical lectures (40%). To ensure accessibility and long-term utility, all teaching materials, datasets, and code repositories were made openly available on GitHub. A dedicated Discord server was established to facilitate asynchronous technical support and peer-to-peer communication. All training sessions were recorded and publicly archived on the project's YouTube channel to serve as a persistent educational resource.

In each year of the project, educational content was themed to introduce emerging technologies, for example 2024 training involved utilizing web-based tools for hydrologic analysis. Similarly, 2025 involved running machine learning models via Google's Collab, and leveraging dedicated high performance computing resources provided by the US National Science Foundation's (NSF) National AI Research Resource (NAIRR) program.

Week 1 (Open Training): The start of each event consisted of theoretical and practical training that was open to all the participants who applied to the program. An orientation was conducted in the first hour of day 1 for the WaterSoftHack Fellows with general introduction of the program as well as code of conduct and expectations. The first half of each day focused on scientific theory in formal lectures, while the second half was structured as hands-on practical training with code implementation demonstration, model development and data processing with real datasets. The final day of the week was restricted to the Fellows and featured an interactive science communication seminar from Alan Alda Center at Stony Brook University. This provided essential training to improve messaging and facilitate team cohesion, serving as a preparatory exercise for team formation ahead of the capstone projects.

Week 2 (Capstone Projects): This week was limited to U.S.-based researchers and included a subset of week 1 participants selected as WaterSoftHack Fellows. Fellows formed interdisciplinary research teams by aligning their expertise and research interests around a central project idea. Program organizers ensured that each capstone team included a balance of junior and senior researchers and maintained a 50/50 gender balance. Teams then developed projects that applied the techniques covered during the training to extract actionable insights from real-world datasets. All activities were conducted under the supervision of faculty mentors and technical trainers. The week concluded with team presentations, a moderated Q&A session, and submission of a short position paper. Projects were evaluated by program judges using a standardized rubric emphasizing technical rigor, reproducibility, and potential scientific and societal impact, with the top-scoring project receiving an award.

3. Evaluation Design and Methods

The development and evaluation of WaterSoftHack cybertraining adhered to a design-based research (DBR) methodology (Design-Based Research Collective, 2003), emphasizing iterative refinement based on empirical evidence and participant feedback. A preliminary need assessment survey was conducted through CUAHSI within the water science community to identify specific training needs and priorities, which

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

directly informed curriculum design. Instructional materials were collaboratively peer-reviewed and made publicly available on GitHub prior to the cybertraining, ensuring transparency and promoting community engagement.

All procedures conformed to ethical research standards. Participants provided informed consent. All data were de-identified and securely stored, and the study obtained Institutional Review Board (IRB) determination prior to analysis. Recognizing limitations such as small cohort sizes and potential self-selection bias, the analysis incorporated triangulation of multiple data sources and emphasized effect sizes, and non-parametric methods to ensure robustness and interpretive rigor. This evaluation framework provides a systematic and ethically grounded approach for assessing program effectiveness, emphasizing both accountability and continuous improvement. The following section presents the results derived from this evaluation, including outcomes of the needs assessment, participant learning gains, and thematic insights from post-cybertraining feedback.

4. Results

4.1. WaterSoftHack Fellow Selections

The selection process for the fellowship program was highly competitive. In its inaugural year, 2024, a cohort of 15 fellows was selected from a pool of 109 US-based applicants. The program continued in 2025 with 130 US-based applications, from which 12 new fellows were chosen. Six of the Fellows from the 2024 cohort continued the WaterSoftHack program in 2025, resulting in a total of 18 fellows for the 2025 program. WaterSoftHack attracted a multidisciplinary group of participants with academic backgrounds spanning multiple disciplines including civil engineering-water resources, environmental engineering, geology, earth science, and computer science. The cohort was intentionally composed of individuals with varying levels of expertise, encompassing both newcomers to the field and those with prior exposure to data science methodologies.

Financial support for the selected Fellows was provided through the NSF CyberTraining program. This funding was instrumental in ensuring the program's accessibility and facilitating broad participation from a diverse pool of candidates. The selection criteria placed a strategic emphasis on fostering a diverse and inclusive environment. Priority was given to applicants from institutions with low cyberinfrastructure adaptation, with a particular focus on those from R2, HBCU, and PUI institutions. The final cohort was deliberately structured to include a representative mix of academic positions, including undergraduate students, master's graduate students, postdoctoral scholars, and early-career faculty. Furthermore, the selection process was committed to enhancing diversity across multiple dimensions, including gender, ethnicity, race, and disability status. A comprehensive demographic breakdown of the participants detailing their institutional affiliations, academic positions, ethnicity, disability, gender, race, and U.S. citizenship status is presented in Figures 1 and 2 for years 2024 and 2025, respectively. As illustrated in Figure 1, the program attracted 33.3% Asian, 26.7% of Hispanic, 20% African, 13.3% Caucasian, and 6.7% North African. The ethnicities of participants in 2025 include 27.8% Caucasian, 22.2% Asian, 16.7% African, 11.1% North African, 11.1% Asian American, 5.6% Hispanic, and 5.6% Native American.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

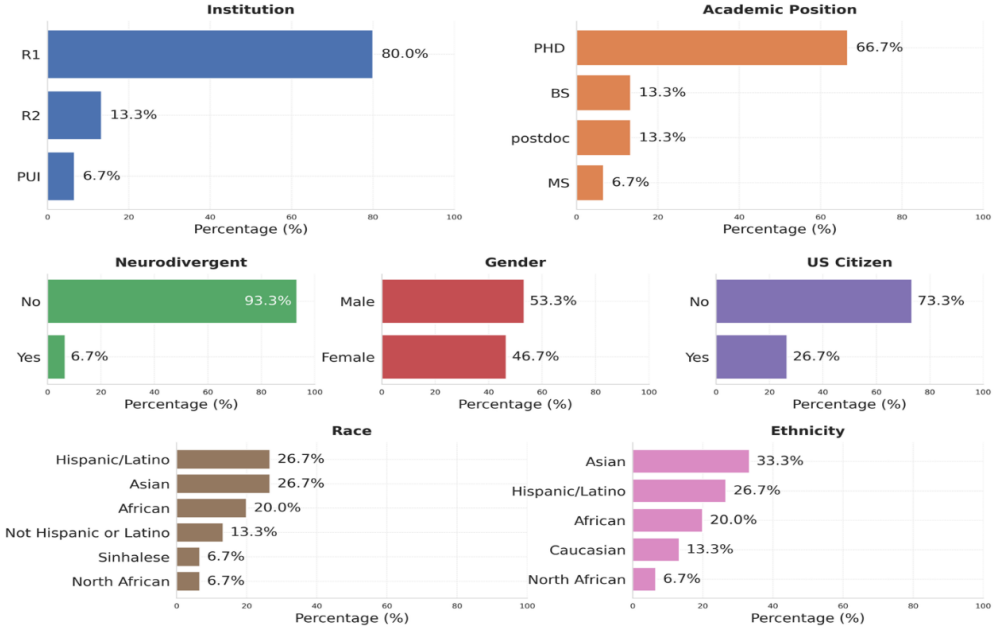

Figure 1. Breakdown of participant details for the WaterSoftHack Fellows that were selected (n=15) for the 2024 cybertraining.

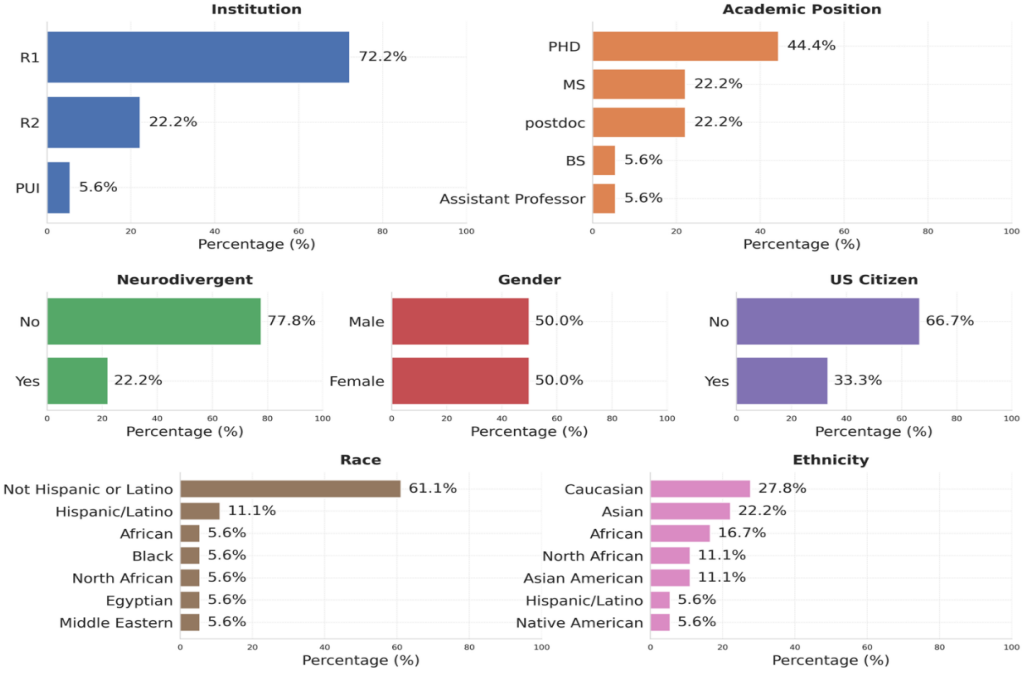

Figure 2. Breakdown of participation details for the WaterSoftHack Fellows selected (n=18) for the 2025 cybertraining.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

4.2. Need Assessment Survey

The WaterSoftHack program was directly informed by a pre-program need assessment survey designed to ensure alignment with workforce and community needs. The survey was distributed through CUAHSI to the broader water science community to systematically identify existing knowledge gaps in data science, machine learning, and cyberinfrastructure applications within hydrology. Findings from the survey guided the selection of training topics, hands-on activities, and capstone project themes, ensuring that the program addressed high-priority skills required for modern hydrologic research and practice. The survey results indicated a significant need for training in fundamental cyberinfrastructure and machine learning literacy and the development of reproducible computational workflows. The findings from 2024 and 2025 need assessments are summarized in Figures 3 and 4, respectively. The 2024 survey (n = 13) focused on identifying skill gaps in data analysis and visualization, revealing strong demand for training in statistical analysis, data visualization, big data technologies, time series analysis, and geospatial data processing across both academic and industry sectors. Similarly, the 2025 survey (n = 30) examined gaps in machine learning competencies, highlighting high demand for skills in deep neural networks, hyperparameter tuning, training efficiency, performance optimization, data parsing, and web-based scraping tool design and implementation. Statistical analysis and data visualization were identified as the most important training skills in 2024 while data parsing and machine learning training efficiency and performance optimization were viewed as the most critical skills to learn in 2025.

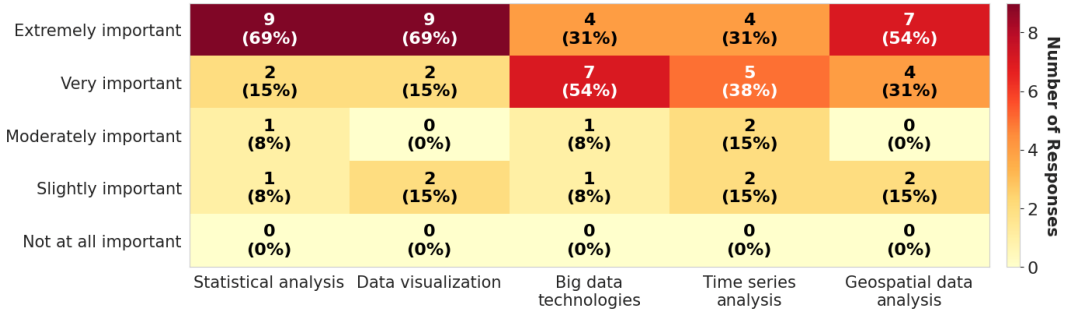

Figure 3. Training need survey results for 2024 with 13 participants completing the survey.

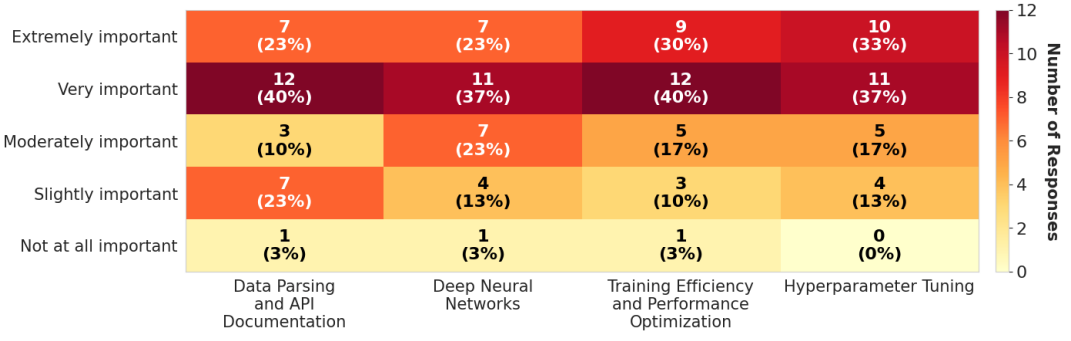

Figure 4. Training need survey results for 2025 with 30 participants completing pre-event survey.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

The WaterSoftHack cybertraining materials and curriculum evolved around three core cybertraining components, data analytics and visualization, machine learning, and cloud/edge computing which were implemented as modular WaterSoftHack training packages and assembled into the WaterSoft Python package. WaterSoft package is hosted as a public monorepository on GitHub, containing training modules, source codes, related datasets, and capstone projects developed by WaterSoftHack Fellows. The details of these training components and the structure of the monorepository are discussed in the following sections.

4.3. WaterSoft Python Package

The preparatory materials provided the foundational knowledge necessary for participants to engage effectively in cybertraining, covering core programming concepts and data science techniques. As of February 2026, the WaterSoft monorepository on GitHub hosts a comprehensive suite of resources, including HydroSuite (Hydrology Software Suite) training modules, data visualization tools, machine learning workflows, hyperparameter tuning models, and capstone projects developed by the WaterSoftHack Fellows. These components collectively form the foundation of the WaterSoft training framework, designed to promote reproducible, data-driven hydrologic research. By the conclusion of the next WaterSoftHack cycle (summer 2026), a dedicated cloud/edge computing training module will be integrated into the WaterSoft package, further expanding its scope. Figure 5 presents a block diagram illustrating the key components and interconnections within the WaterSoft framework. This includes three different types of workflows: data analytics and visualization, machine learning, and cloud and edge computing.

Beyond its computational training components, the WaterSoft Python package also incorporates modules that directly support hydrological science applications. The package provides tools and workflows for accessing and processing hydrologic and meteorological datasets, performing time-series analysis of hydrological observations, and developing simulation models for hydrologic variables such as streamflow, river stage, and water-quality indicators. For example, training modules within the HydroSuite framework demonstrate how machine learning and statistical methods can be applied to hydrological forecasting problems using observational datasets from sources such as stream gauges and meteorological stations. These workflows include data acquisition through APIs, preprocessing and quality control of hydrological time-series data, feature engineering for watershed variables, hyperparameter tuning, and the development of predictive models for hydrologic processes.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

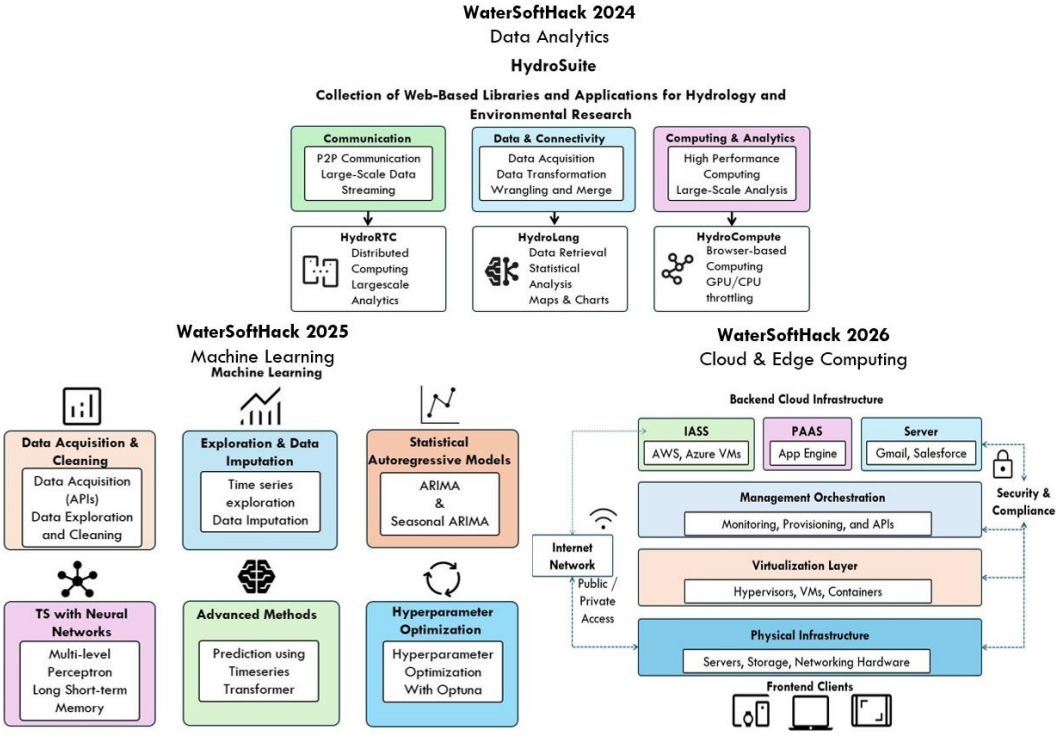

Figure 5. The workflow of the WaterSoft Python package which integrates three primary components - data analytics and visualization, machine learning, and cloud and edge computing to enable scalable, data-driven water science research and training.

4.3.1. HydroSuite Training Module

Year 1 (2024) of WaterSoftHack involved a comprehensive introduction to HydroSuite and HydroLang (Hydrology + Language; Ramirez et al., 2022), HydroCompute (Hydrology + Compute; Ramirez et al., 2024a), HydroRTC (Hydrology+Real-Time Communication; Ramirez et al., 2024b), and HydroSuite AI Helper (Pursnani et al., 2024). The pedagogical approach for the module was predicated on the principles of accessibility and immediate applicability. To remove the need for complex local software configuration, the instruction was delivered via online, open-access sandbox environments. This methodology ensured a standardized and universally accessible platform, allowing participants to execute and modify code directly within a web browser, thereby focusing on data analytics concepts rather than environment setup. HydroSuite contained three different modules: The initial phase of the training centered on HydroSuite, a JavaScript framework for client-side data acquisition, analysis, and visualization. The curriculum was then continued around computing four other modules including HydroLang, HydroCompute, HydroRTC, and HydroSuite AI Helper. Data module component instructed participants on programmatic interaction with environmental data. The tutorials covered methods for data retrieval from diverse sources with public application programming interfaces (APIs) and model repositories. Participants learned techniques for data manipulation and transformation, as well as client-side data input/output operations.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

Additionally, the training introduced multiple computational capabilities of HydroLang including a hydro component containing functions for fundamental hydrological analyses like rainfall-runoff modeling; a stats component for statistical characterization of datasets; and a neural network component, which demonstrated the implementation of feed-forward neural networks using the TensorFlow.js library for in-browser machine learning applications.

Visualization and maps modules facilitated the interpretation and dissemination of results, with subsequent training focused on data presentation. The visualization module provided instructions on rendering dynamic charts and tables from analytical outputs using the Google Charts library. The maps module equipped participants with the skills to generate interactive geospatial visualizations, including the rendering of GeoJSON and Keyhole Markup Language (KML) data layers, utilizing either the Leaflet or Google Maps platforms.

The second section of the training module addressed the challenge of computationally intensive tasks through the HydroCompute library (Ramirez et al., 2024a). This library is engineered to execute high-performance simulations on the client side, thereby leveraging the computational resources of the end-user device. The instructions detailed HydroCompute's multi-engine architecture, which supports computations via native JavaScript, WebAssembly (WASM), and WebGPU. The inclusion of WASM is particularly significant, as it allows the execution of code compiled from high-performance languages like C and C++ at near-native speeds. The WebGPU engine allows for the utilization of a device's Graphics Processing Unit (GPU) for parallelizable scientific computations. The training workflow involved data loading, defining computational sequence with specified functions and dependencies, and the execution of the simulation, demonstrating the feasibility of conducting sophisticated modeling within a standard browser environment.

The final component of the training introduced HydroRTC, a library designed to facilitate data sharing and collaborative analysis. Built upon WebRTC and Socket.IO, HydroRTC provides a framework for both server-to-peer (S2P) and peer-to-peer (P2P) communication. The curriculum explored several advanced use cases. Participants learned to implement S2P data streaming for the efficient distribution of large datasets, such as remote sensing imagery. A key concept presented was "smart data transmission," an approach where an initial analysis on a data subset informs the prioritization of subsequent data delivery.

Furthermore, the training provided in-depth instruction on P2P architecture. This paradigm enables the direct exchange of data and model results between clients, which can significantly reduce server load and create decentralized, collaborative environments. The practical applications demonstrated included volunteer computing for distributed modeling tasks and real-time data sharing among research collaborators which highlight a shift towards more interactive web-native scientific workflows.

4.3.1.1. Case Study

During the training session, several case studies were presented as part of the accompanying session materials, with the primary focus centered on demonstrating the interoperability and connectivity within the HydroSuite collection of libraries spanning data acquisition, wrangling, and sorting through advanced visualization for extracting insights from hydrological datasets. A representative case study examined the Salt Lake City, Utah, USA using U.S. Geological Survey (USGS) gaging data, within which ARIMA (Autoregressive Integrated Moving Average), Seasonal ARIMA (SARIMA), and Markov Chain Monte Carlo modeling processes were implemented to quantify predictive trends derived from raw or minimally processed data. The computations were performed dynamically on the user's local machine through a web-

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

based containerized environment, enabling participants to explore the underlying mechanisms inherent in time-series hydrological data and to assess model responsiveness to real-world variability. By employing web technologies and client-side hardware, the session further exposed participants to contemporary, browser-based computational approaches that facilitate rapid prototyping and model deployment beyond conventional development environments. These technologies supported the creation of interactive dashboards and analytical utilities, allowing end users to conduct independent analyses across selected geographic regions. This integration is exemplified in Figure 6, which presents the predictive outputs and spatially contextualized information overlaid on a geospatial map illustrating the associated datasets.

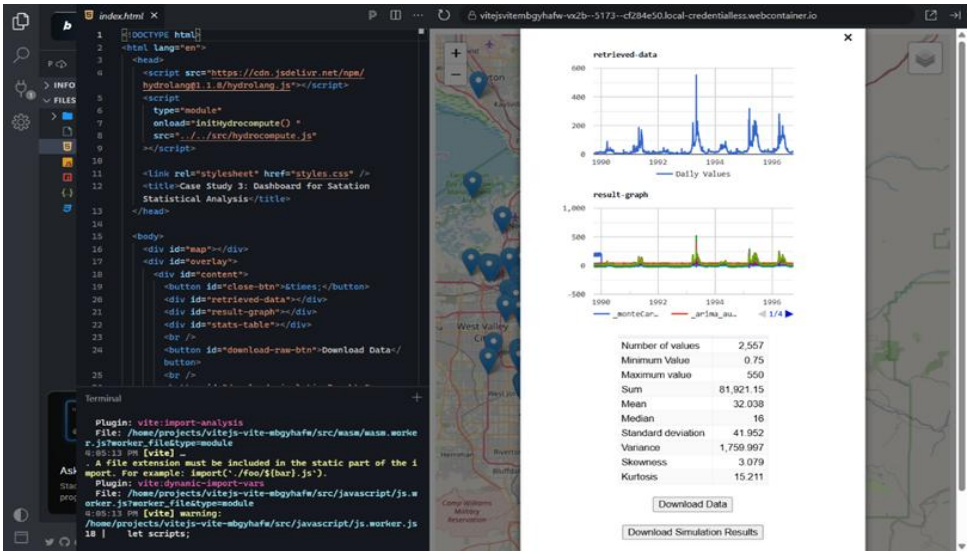

Figure 6. Containerized app code shared with the participants in the 2024 WaterSoftHack cybertraining for analyzing time series datasets obtained from multiple USGS stations.

4.3.2. Machine Learning Training Modules

Year 2 of WaterSoftHack cybertraining was dedicated to machine learning applications which provided participants with a structured curriculum and hand-on training that progressed from foundational data handling to advanced deep learning applications. The modules were designed to equip attendees with the theoretical knowledge and practical skills necessary to implement machine learning solutions for complex hydrological challenges. The initial phase of the training focused on the complete workflow development of time series analysis, beginning with data acquisition, preparation, and normalization. Participants were instructed on the programmatic retrieval of hydrological and meteorological datasets using APIs.

Following acquisition, the curriculum addressed critical data preprocessing techniques, including methodologies for the imputation of missing values and the identification and treatment of outliers that are frequently encountered in water science data. A significant component was dedicated to feature engineering, wherein participants learned how to derive and select influential predictor variables from raw time series data to enhance model performance. Upon establishing a foundation in data preparation, the module introduced classical statistical methods for regression prediction. Instruction covered the theoretical

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

underpinnings and practical application of ARIMA and SARIMA models. These sessions provided participants with a robust baseline for understanding and modeling temporal dependencies and seasonality in hydrological data.

Building upon the statistical methods, the training program progressed to the application of neural networks for water science time series prediction. This section began with an introduction to the Multi-Layer Perceptron (MLP), a foundational class of feedforward artificial neural networks. The curriculum then advanced to more complex recurrent architectures, with a specific focus on Long Short-Term Memory (LSTM) networks. The training emphasized the advantages of LSTM in capturing long-term temporal dependencies, a critical characteristic for many hydrological forecasting tasks.

The final phase of the WaterSoftHack week 1 training exposed participants to state-of-the-art machine learning techniques and their real-world applications. Instructions were provided on the use of Transformer models for time series forecasting which showed the architecture's powerful attention mechanism for modeling complex sequences in time series data. To ensure the development of high-performing and generalizable models, the curriculum included a dedicated segment on systematic hyperparameter optimization methods as well.

4.3.2.1. Case Study

To consolidate the theoretical instruction, a parallel hands-on computation case study was demonstrated to the participants. This case study was designed to provide practical experience in applying statistical and machine learning models to real-world hydrological prediction tasks. The objective was to predict river gauge height in a USGS gauging station one hour in advance. The models utilized a multivariate time series dataset from the USGS gauging station 02336490. Hourly data from 2008 to 2024 was used for training, testing and validation. This station is located at the outlet of the Proctor Creek-Chattahoochee River watershed (HUC 031300020101) in Fulton County, Georgia (33°49'02.7" N, 84°28'49.2" W, NAD83). Complementary meteorological data which included precipitation, air temperature, wind speed, relative humidity and air pressure were obtained from the proximate Fulton County Airport monitoring station (station id 03888). All data, including gauge height and meteorological variables, were aggregated to an hourly temporal resolution.

Participants were guided through a complete data science workflow using computational notebooks provided via Google Colaboratory. The standardized workflow included scripts for data acquisition, quality control, imputation of missing values, and feature engineering. The modeling task was formulated as predicting the gauge height at time $t + 1$ using a lookback window of the preceding 24 hours of observations (both gauge height and meteorological covariates). The complete dataset was chronologically partitioned into training (70%), validation (10%), and testing (20%) subsets to prevent temporal data leakage. Five distinct time series forecasting models were developed and compared: ARIMA, SARIMA, MLP, LSTM, a Transformer architecture. For the three machine learning models (MLP, LSTM, and Transformer), hyperparameter optimization was demonstrated using the Optuna optimization framework (Akiba et al., 2019). For the ARIMA model, the autoregressive order (p), differencing order (d), and moving-average order (q) were set to 5, 1, and 5, respectively. These values were selected empirically through trial-and-error, considering both predictive accuracy and computational cost. For the SARIMA model, the same non-seasonal orders $(p, d, q) = (5,1,5)$ were used, together with seasonal orders

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

$(P, D, Q) = (2,0,2)$ and a seasonal period of 7. Relatively small seasonal orders were chosen to keep model training computationally feasible within the available time budget. The optimization process was configured to execute 100 trials, with the objective of minimizing the validation loss. Parameters were sampled from the specified intervals using t ree-structured Parzen Estimator (TPE) (Bergstra et al., 2011). The final optimized hyperparameters obtained for each architecture are reported in Table 1.

As shown in Table 1, the learning rate, which controls the step size during training, is highest for the LSTM (0.00103), lowest for the MLP (0.00022), and intermediate for the Transformer (0.00059). This reflects differences in how quickly each model adjusts its weight. Weight decay, which is a regularization parameter that penalizes large weights to reduce overfitting, is strongest for the LSTM (0.00025) while it is weakest for the MLP (3.22×10^{-5}) and moderate for the Transformer (0.00016). Dropout, another form of regularization where neurons are randomly deactivated during training, is not used in the LSTM, but it was set at 0.19 for the MLP, and slightly higher at 0.195 for the Transformer. This reflects the fact that the Transformer applies a smaller dropout of 0.110 to positional encoding to regularize the attention mechanism. Overall, these hyperparameter choices show that the LSTM prioritizes faster learning with strong weight decay while the MLP relies on careful weight updates with dropout for regularization, and the Transformer balances learning rate and weight decay while incorporating multiple layers of regularization to handle complex dependencies.

Table 1. Optimized hyperparameters for three different machine learning models demonstrated in the case study (Only a subset of hyperparameters were selected for demonstration).

Hyperparameters	LSTM	MLP	Transformer
Learning Rate	0.00103	0.00022	0.00059
Weight Decay	0.00025	3.22×10^{-5}	0.00016
Dropout	-	0.19	0.195
Dropout (Positional encoder)	-	-	0.110

Model performance was quantitatively evaluated on the unseen test dataset using two standard hydrological metrics: the Nash-Sutcliffe Efficiency (NSE) and the Mean Absolute Error (MAE). This practical application enabled a direct comparison of the models' respective abilities to capture complex temporal dynamics. Figure 7 provides a visual comparison of the predicted versus observed gauge heights for all five models over a representative one-week period for USGS02336490 gauging station. The USGS gauging station 02336490 is situated on the main stem of the Chattahoochee River near Atlanta, Georgia, downstream of several major reservoirs and receiving inflows from urbanized tributaries such as Proctor Creek. As illustrated, the ARIMA and SARIMA models achieved strong predictive performance which reflect the relatively stable and autocorrelated flow regime of this reservoir-regulated river segment. The LSTM model also demonstrated competitive skill. In contrast, the Transformer model exhibited lower predictive accuracy and tended to underestimate peak flows This suggests that attention-based architecture may require larger datasets, inclusion of additional hydrological and meteorological covariates, or tailored training strategies to effectively capture the complex dynamics of gauge height at this urban-influenced and regulated site.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

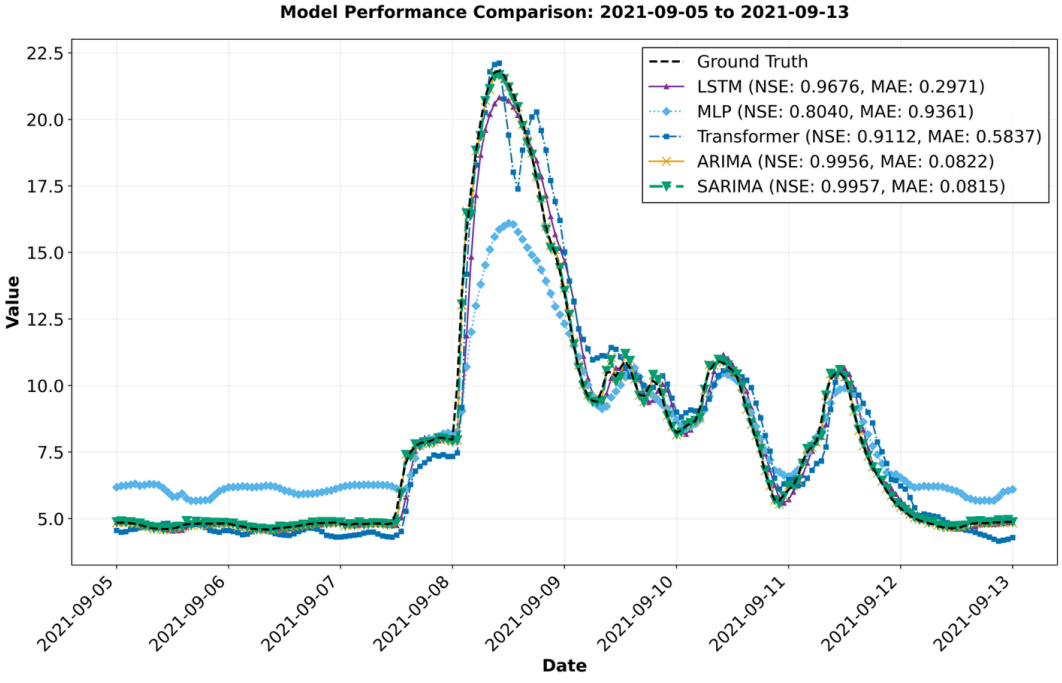

Figure 7. Comparison of predicted and observed gauge height data for September 5-13, 2021 in the USGS gauging station 02336490. We used five models including ARIMA, SARIMA, LSTM, MLP and Transformer to calibrate streamflow. The plot also displays the NSE and MAE values calculated for each model over the test dataset.

4.3.3. Cloud and Edge Computing Module

The planned Year 3 module will introduce participants to cloud and edge computing frameworks and their applications in water science and engineering. Building on knowledge acquired during the first two years, the module will aim to familiarize participants with various cloud and edge computing paradigms, commercial platforms, and water science–specific tasks that can be addressed using these technologies.

The training will begin with the fundamentals of cloud and edge computing. It will introduce the participants to the concepts, architecture, and technologies that enable modern distributed systems. The training will cover core cloud computing models such as Infrastructure as a Service (IaaS), Platform as a Service, and Software as a Service, along with key characteristics like scalability, virtualization, and on-demand resource provisioning. It will also cover edge computing fundamentals, focusing on processing data closer to the source to reduce latency, to improve real-time performance, and to optimize bandwidth usage. Finally, participants will see how to leverage cloud and edge computing for water science applications.

Participants will learn to access and navigate the cloud and edge platforms via Cloudbank, with IBM Cloud and its Environmental Intelligence Suite serving as the primary training platform. This suite provides ready-to-use APIs and software development kits for a wide range of applications, including geospatial queries,

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

weather analytics, greenhouse gas emissions calculations, climate risk assessment, high-resolution imagery, and large curated datasets—making it particularly well-suited for water science projects. The curriculum will also cover optimal storage strategies for large hydrological datasets. The module will conclude with a comprehensive case study, where participants will apply these tools to a real-world project.

4.4. WaterSoftHack Training Evaluation

Pre- and post-surveys were conducted with WaterSoftHack Fellows to assess the effectiveness of the training. The evaluation followed a design-based research methodology, emphasizing iterative refinement of WaterSoftHack materials based on participant feedback. This approach enabled continuous improvements to both instructional content and delivery methods, informed by data collected during the two years of implementation. Data were collected by the educational research team through pre- and post-cybertraining surveys as well as follow-up online interviews with Fellows. The surveys included Likert-scale (Likert, 1932) and open-ended questions designed to evaluate satisfaction, skill development, and confidence in applying the training to participants' research projects. Follow-up interviews were semi-structured, focusing on participants' overall experiences, perceived challenges, and recommendations for enhancing the program.

The pre- and post-WaterSoftHack survey responses varied between years. In 2024, 13 of 15 Fellows completed the pre-cybertraining survey, and 12 responded to the post-cybertraining survey (Figure 8). In 2025, 11 of 18 Fellows completed the pre-survey, while only 8 responded to the post-survey (Figure 9). Analysis of the raw survey data indicates a reduction in the proportion of participants who self-identified as "novices" across nearly all skill categories (see Tables 2 and 3), apart from software development in 2025, which is planned to be explicitly addressed in the 2026 training. Figures 8 and 9 illustrate a clear decline in novice-level self-assessments for both years. For instance, in 2025, only 45.5% of participants rated themselves as competent or proficient in machine learning before the training, whereas this proportion increased to 87.5% afterward. Similarly, in 2024, only 23.1% of participants considered themselves above novice level in data analytics and prototyping before the WaterSoftHack cybertraining, but it increased to >50% after the completion. These trends suggest notable self-reported improvements across multiple skill domains. Formal statistical analysis of these changes was not possible because the post-training survey data were anonymous, unmatched, and limited in number. As a result, the dataset does not allow for a quantitative estimate of the program's impact on skill development. The evaluation of WaterSoftHack's effectiveness therefore draws primarily on qualitative evidence.

Table 2. Comparison of self-reported novice counts from the 2024 pre- and post WaterSoftHack-cybertraining surveys. "Expected Novices" is the number that would be expected in the post-survey if the proportion of novices had remained unchanged from the pre-survey.

Skill Category	Observed Novices (after cybertraining)	Expected Novices considering no improvement (%)
Software Development	6	9.23
Edge Computing	1	3.69
Cloud Computing	6	8.31
Machine Learning Programming	8	10.15
CI Prototyping	8	9.23

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

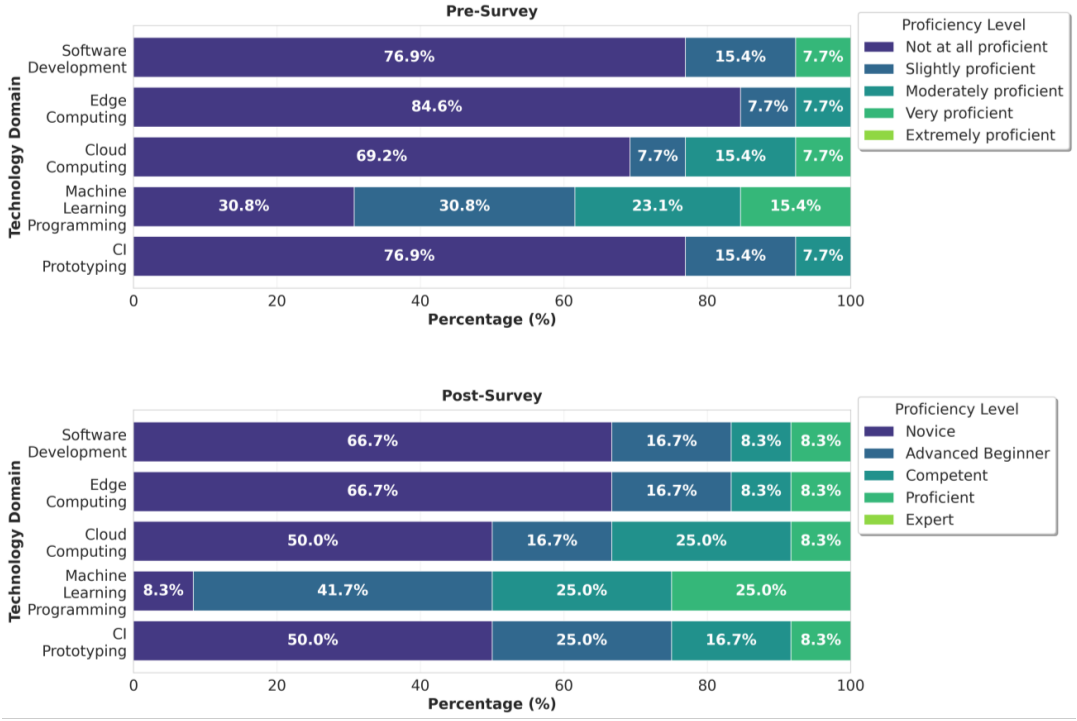

Figure 8. The results of pre- and post-cybertraining survey results conducted among the Fellows for WaterSoftHack 2024. 13 out of 15 Fellows responded to the pre-survey and 12 out of 15 Fellows responded to the post-survey.

Table 3. Comparison of self-reported novice counts from the 2025 pre- and post- cybertraining surveys. "Expected Novices" is the number that would be expected in the post-survey if the proportion of novices had remained unchanged from the pre-survey.

Skill Category	Observed Novices (after cybertraining)	Expected Novices considering no improvement (%)
Software Development	6	5.81
Edge Computing	5	6.54
Cloud Computing	2	4.36
Machine Learning Programming	1	2.18
CI Prototyping	2	6.54

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

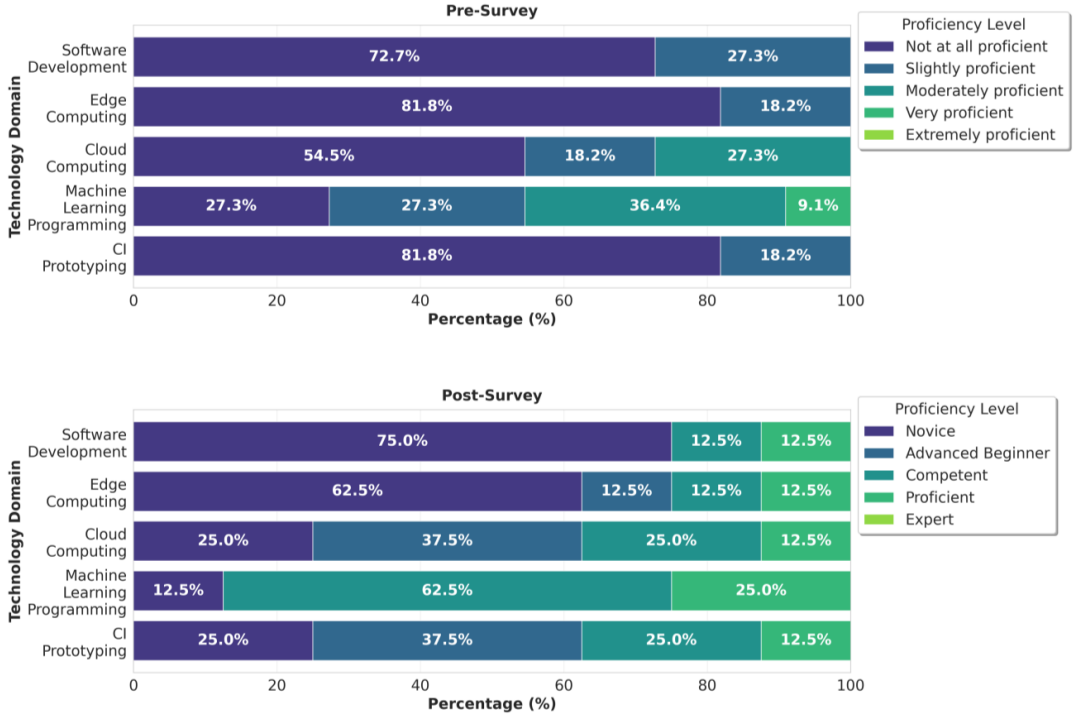

Figure 9. The results of pre- and post-WaterSoftHack cybertraining survey results conducted among the Fellows for WaterSoftHack 2025. 11 out of 18 Fellows responded to the pre-survey and 8 out of 18 Fellows responded to the post-survey.

4.5. Participant Experience

Post-event surveys from the 2024 and 2025 editions of WaterSoftHack cybertraining consistently revealed strong positive feedback regarding networking and collaboration among participants. Throughout both years, attendees praised the event for providing valuable opportunities to connect with peers in related fields, which in turn fostered new research partnerships spanning various institutions and disciplines. Participants highlighted that these connections often extended beyond the event itself, with many engaging in ongoing communication or team projects started during the program.

A particularly notable strength of WaterSoftHack, as identified in participant feedback, was the practical focus of its hands-on components. The majority of respondents reported that these sessions helped them apply and learn computational techniques, such as machine learning and advanced cyberinfrastructure workflows, which they planned to integrate into future research and teaching. These practical sessions were frequently mentioned as the most useful and engaging, giving participants the confidence and skills needed to apply new methods in their work.

Survey responses further indicated that the program's collaborative environment contributed to tangible outcomes. Several participants from both years stated specific intentions to develop their hackathon or capstone projects into publishable studies. Some were already collaborating on joint publications or follow-on research initiatives spurred by connections made during the event. The structure of the program,

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

emphasizing interdisciplinary teamwork and accessible mentorship, was credited with expanding participants' research capacity and confidence in using new computational tools. Statistical outcomes support these themes in 2024 and 2025 post-survey results.

WaterSoftHack post-event survey indicated that the majority of respondents recognized the program as highly effective in connecting traditional water science with modern computational methodologies, with 60% rating it as "Very Effective" or "Extremely Effective". Participants consistently reported expanded knowledge in applying machine learning to water science, including gaining foundational theoretical understanding, enhancing practical coding abilities, and successfully integrating new methods into authentic research problems such as streamflow prediction and data-driven hydrologic modeling.

Survey results revealed that open-source machine learning models and collaborative tools led 59% of participants to significantly change their approach to computing in water science, notably improving their predictive capacities and supporting real-time, cloud-enabled collaborative modeling efforts. Over half of the respondents reported increased competence in building and using advanced cyberinfrastructure workflows utilizing platforms like Google Colab, NAIRR, and GitHub, which were praised for their accessibility and direct relevance to research projects.

Participants highlighted several key strengths of WaterSoftHack cybertraining program. Major strength among them was the hands-on, applied learning environment, which combined with expert mentorship, created a powerful setting for skill development and confidence-building. Most respondents indicated that the sessions and tools learned were immediately transferable to their work, either supporting ongoing research or facilitating future teaching and outreach. Collaboration emerged as a major benefit, with many initiating new partnerships, sharing resources, and launching joint projects during and after the program.

While satisfaction and session ratings were overwhelmingly positive, especially for interactive, project-based learning that reinforced concepts and skills, participants also identified opportunities for future improvement. These included extending the program duration beyond two weeks, providing more discipline-specific case studies, improving team matching, and clarifying educational materials to increase inclusiveness for diverse interdisciplinary backgrounds. Furthermore, several participants commented on scheduling, noting that the 8:00 a.m. Pacific Time start was challenging for those on the US West Coast, and recommended more flexible timing to accommodate a wider geographic audience.

Feedback also reflected a strong appreciation for the program's collaborative and supportive culture, impactful technical content, and opportunities for networking with peers and mentors. Overall, participants described substantial professional growth and valued the chance to develop new skills, apply them in research, and form lasting collaborations. The program's applied and collaborative approach was seen as a model for advanced training at the intersection of water science and computational methods. Comparing results from both 2024 and 2025, the WaterSoftHack has consistently satisfied key needs of the water science community. While some 2024 participants noted that late delivery of training materials limited their preparation, this issue was addressed in 2025 by providing resources six months in advance which markedly improved pre-training readiness. Many also recommended earlier team formation and project initiation during week one to boost collaboration efficiency, a suggestion that is being considered for implementation in the 2026 WaterSoftHack iteration.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

Findings from both years' surveys consistently underscore WaterSoftHack's significant impact on professional development, collaborative research, and the application of computational methods in water science. This shows the program's ongoing value as a model for interdisciplinary, cybertraining-driven scientific capacity building for water science and engineering.

4.6. Qualitative Insights from WaterSoftHack Fellows

CUAHSI conducted post-event reflections with WaterSoftHack Fellows to illustrate the program's influence on research practices and professional development (CUAHSI, 2025; 2026). One doctoral participant reported that training in HydroLang significantly enhanced their ability to process and visualize large remote-sensing datasets. These newly acquired skills are now being applied in their dissertation research on water quality monitoring. The participant described plans to develop a web-based API for automated monitoring using techniques learned during the workshop. Another fellow emphasized the networking opportunities facilitated by the program, noting that connections formed during WaterSoftHack evolved into ongoing professional relationships reinforced at subsequent scientific meetings.

A postdoctoral scholar highlighted the value of centralized and curated training resources that reduced the fragmentation commonly encountered in self-directed learning. Collaborative work during the hackathon resulted in the creation of shared resources on HydroShare and a continental-scale assessment of stream temperature variability, which was later presented at a major scientific conference and is currently being prepared for journal submission (CUAHSI, 2025). Reflections from the 2025 fellows further reinforced these themes. Participants reported that hands-on machine learning modules covering model development, hyperparameter tuning, and evaluation directly improved their ability to build predictive models for environmental applications such as sediment forecasting and water quality monitoring (CUAHSI, 2026). Fellows also highlighted the importance of interdisciplinary teamwork during capstone projects, which enabled participants from different institutions and disciplinary backgrounds to collaborate on real-world water science challenges and expand their professional networks.

Several fellows indicated that the skills gained during WaterSoftHack were immediately transferable to their graduate research activities. Examples included developing satellite-based models for suspended sediment estimation using machine learning, applying LSTM models to downscale satellite hydrologic datasets, and experimenting with ensemble machine learning approaches for environmental forecasting (CUAHSI 2025, 2026). These reflections collectively illustrate how the program not only builds technical competencies but also fosters collaborative research environments that improve the adoption of data-driven methods in water science.

4.7. Candidate Follow-up and Career Outcomes

Across the two candidate cohorts (2024–2025), a total of 27 WaterSoftHack Fellows were recruited into the program, with each cohort comprising 15–18 candidates. In addition to the Fellows, the program provided training to over 350 participants, of whom approximately 40% identified as female and 60% as male. Due to the constraints of NSF funding, participation in the WaterSoftHack Fellowships was limited to U.S.-based candidates (100%). However, the general training (week 1) attracted a diverse international audience, including participants from Europe, Asia, the Middle East, and Africa. Most candidates entered the program after completing a postgraduate master's degree, while a smaller subset had completed a 4-year undergraduate degree or a bachelor's degree. Consistent with the program's interdisciplinary mission,

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

candidates were drawn from a wide range of water science and engineering disciplines, as illustrated by the subject-area word cloud in Figure 10.

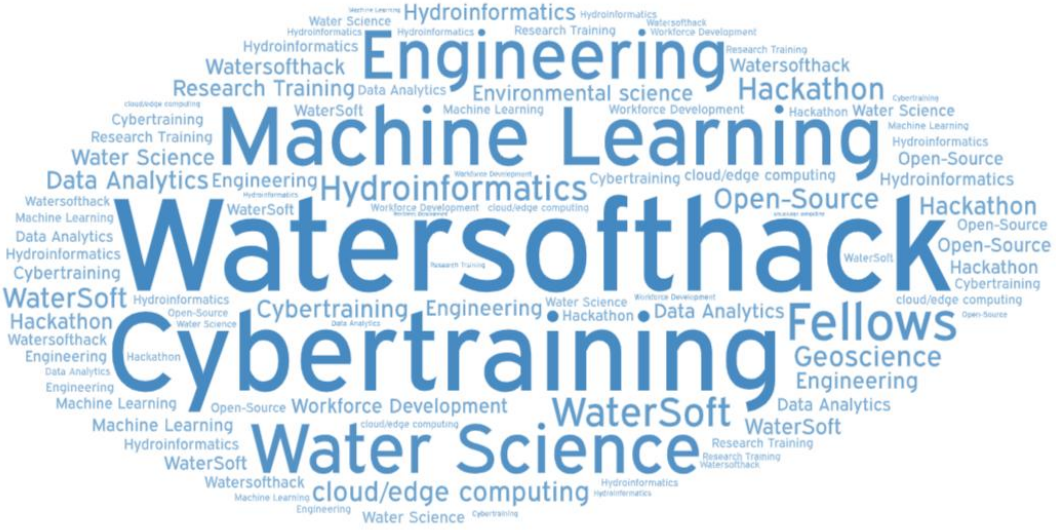

Figure 10. Word cloud summarizing the subject areas covered in WaterSoftHack training during the 2024–2025 program cycles.

Given that only two years of participants have graduated to date (primarily from the two cohort), career outcome assessment remains preliminary. Nonetheless, experience from similar training programs indicates that strong research interactions during training can meaningfully influence long-term career trajectories (Wagener et al., 2021). Among the WaterSoftHack Fellows, 3 of 27 have already secured tenure-track faculty positions in the US, and several others have obtained postdoctoral appointments. In addition, the WaterSoft package developed through this program is now being incorporated into teaching curricula at multiple institutions. For example, it is currently used in CEE5310: Informatics for Sustainable Systems at the University of Iowa and AGM8710: Computational Methods in Water Resources at Clemson University.

5. Discussion and Conclusion

The rapid growth of large environmental datasets and advanced computational tools has exposed a widening gap between traditional water science education and the data-driven skills required to address modern challenges in hydrology. Recognizing this need, the WaterSoftHack cybertraining program was designed to directly immerse students and early-career researchers in a collaborative learning environment focused on data science, machine learning, and cloud and edge-enabled scientific workflows. Through hands-on training and project-based learning, participants gained practical experience that bridged theoretical knowledge with modern computational approaches while it also fostered collaboration and network building within the water science community.

Program evaluation relied primarily on qualitative feedback, given certain constraints in the quantitative survey design. Nevertheless, participant responses overwhelmingly indicated that WaterSoftHack was highly effective in enabling the integration of new computational methods, promoting interdisciplinary

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

teamwork, and empowering researchers to apply these approaches in their ongoing and future research. Many participants credited the training with advancing their technical skills, research productivity, and openness to innovation.

At the same time, the process of developing and implementing WaterSoftHack revealed important challenges such as the need for longer-duration training, improved preparation and team formation, and more adaptive materials for a diverse audience spanning earth and water sciences. One of the most significant challenges was fostering deep collaboration in a short-term virtual setting. Participants appreciated the program's networking opportunities, collaborative spirit, and supportive environment, which enabled valuable peer connections across institutions and time zones. However, many noted that earlier and more structured team formation could further enhance these connections. The fast pace and intensity of the one-week hackathon offered limited time for teams to develop cohesion, balance different technical backgrounds, and converge on shared research objectives. These observations suggest that allocating additional time for team building and project ideation potentially extending collaborative activities over longer periods, such as an academic semester can further enhance the effectiveness of hands-on cybertraining.

Building on these lessons, future iterations of WaterSoftHack incorporate pre-cybertraining virtual mixers and dedicate deliberate time in the first week for project ideation, team building, and familiarization with collaborative tools. Structured activities to refine research goals and finalize teams before the official start of the hackathon would give participants a critical head start, allowing them to enter the intensive phase with a clear direction and stronger rapport.

Another overarching challenge was evaluating program impact in a meaningful way. As highlighted in the results section, small cohort sizes and the need for anonymous, unmatched survey design limited the ability to derive statistically valid conclusions about improvements in skill and knowledge. This limitation underscores a larger issue facing intensive, short-term training initiatives. The experience emphasized two key lessons: qualitative feedback that was gathered through interviews and open-ended survey responses proves invaluable and should be given prominent consideration. In addition, robust evaluation required both matched pre- and post-surveys as well as longer-term tracking. Future evaluations should therefore incorporate follow-up surveys conducted six to twelve months after the program to better capture long-term outcomes. This approach can reveal the broader impacts, such as successful application of learned skills, the development of new open-source tools, and research publications emerging from program participation.

Integrating computational skills into water science marks more than just a technical enhancement. It signals a transformative paradigm shift in how scientific knowledge within the field of hydrology is generated, interpreted, and translated into decision. By weaving computational competencies into the fabric of water science education, the community stands poised to fully harness today's data-rich environment to unlock the capacity for more sophisticated, predictive insights into critical issues in hydrology.

The WaterSoftHack program demonstrates that intensive, hands-on cybertraining can meaningfully improve the transition toward data-driven hydrologic research. Participants not only gained foundational and advanced cyberinfrastructure expertise but also demonstrated how such skills can immediately impact real-world research, team projects, and lasting collaborations across institutional boundaries. Through mentorship, applied learning, and continual refinement, WaterSoftHack reinforces the importance of moving beyond traditional hydrological training programs to more modular and collaborative learning

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

frameworks as the field evolves. With this vision, WaterSoftHack provides a scalable model for preparing the next generation of water scientists to effectively operate in an increasingly digital and data-rich research environment.

6. Acknowledgements: This work is supported by the U.S. National Science Foundation (NSF) Office of Advanced Cyberinfrastructure (AOR2320979). All opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. NSF NAIRR is acknowledged for its generous allotment of computing time for WaterSoftHack cybertraining. Alan Alda Center at Stony Brook University (USA) is acknowledged for hosting the scientific communication session. We acknowledge the participation of students, researchers and early career faculty members across CUAHSI's >130 university partner networks.

7. Software Availability

Name of package: WaterSoft Python Package

Year first available: 2024

Developers: WaterSoftHack Team, Clemson University, Tulane University, University of Iowa, & CUAHSI

Package Availability: https://github.com/watersofthack/WaterSoft

License: MIT License

Software requirements: Python ≥3.10, numpy, pandas, scipy, matplotlib, seaborn, plotly, tensorflow, torch, scikit-learn, statsmodels, geopandas, folium, requests, optuna, jupyter, notebook, dash, websockets, python-socketio

8. Data availability

The hydrological data used in this study are publicly available from the U.S. Geological Survey National Water Information System (NWIS) at https://waterdata.usgs.gov/nwis/rt. All human-subject–related data referenced in this paper are fully aggregated and are provided within the tables of the manuscript.

9. Author contributions

KP and VS prepared and edited the draft manuscript. All authors contributed to the final version of the manuscript.

10. Ethics Statement

The studies involving human participants were reviewed and approved by Clemson University Institutional Review Board. The participants provided their written informed consent to participate in this study.

11. Competing interests

The authors declare that they have no conflict of interest.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

12. References

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., & Rasmussen, J. (1986). An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 1: History and philosophy of a physically-based, distributed modelling system. Journal of Hydrology, 87(1–2), 45–59.

Abbott, M. B., Bathurst, J., Cunge, J., O'connell, P., & Rasmussen, J. (1986). An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 2: Structure of a physically-based, distributed modelling system. Journal of Hydrology, 87(1–2), 61–77.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Anslow, C., Brosz, J., Maurer, F., & Boyes, M. (2016). Datathons: An experience report of data hackathons for data science education. Proceedings of the 47th ACM Technical Symposium on Computing Science Education, 615–620.

Bergstra, J., Bardenet, R., Bengio, Y. and Kégl, B., 2011. Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24.

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrological Sciences Journal, 24(1), 43–69.

Bjerknes, V. (1904). Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. Meteor. Z., 21, 1–7.

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. Educational Psychologist, 26(3–4), 369–398.

Briscoe, G. (2014). Digital innovation: The hackathon phenomenon.

Brooks-Harris, J. E., & Stock-Ward, S. R. (1999). Workshops: Designing and facilitating experiential learning. Sage Publications.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

Brown, J. D., Wu, L., He, M., Regonda, S., Lee, H., & Seo, D.-J. (2014). Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. Journal of Hydrology, 519, 2869–2889.

Charney, J. G., Fjörtoft, R., & Neumann, J. von. (1950). Numerical integration of the barotropic vorticity equation. Tellus, 2(4), 237–254.

Cikmaz, B. A., Yildirim, E., & Demir, I. (2025). Flood susceptibility mapping using fuzzy analytical hierarchy process for Cedar Rapids, Iowa. International Journal of River Basin Management, 23(1), 1-13.

Creswell, J. W., & Clark, V. L. P. (2017). Designing and conducting mixed methods research. Sage publications.

CUAHSI. 2025. Community guest spotlight with 2024 WaterSoftHack Fellows. Retrieved January 26, 2026 from https://www.cuahsi.org/community/news/april-e-newsletter-community-guest-spotlight-with-2024-watersofthack-fellows

CUAHSI. 2026. Community guest spotlight with 2025 WaterSoftHack Fellows. Retrieved March 13, 2026, from https://www.cuahsi.org/community/news/march-e-newsletter-community-guest-spotlight-with-2025-watersofthack-fellows

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., & Fresch, M. (2014). The science of NOAA's operational hydrologic ensemble forecast service. Bulletin of the American Meteorological Society, 95(1), 79–98.

Demir, I., Xiang, Z., Demiray, B., & Sit, M. (2022). WaterBench-Iowa: a large-scale benchmark dataset for data-driven streamflow forecasting, Earth Syst. Sci. Data, 14, 5605–5616.

Demir, I., Conover, H., Krajewski, W. F., Seo, B.-C., Goska, R., He, Y., McEniry, M. F., Graves, S. J., & Petersen, W. (2015). Data-enabled field experiment planning, management, and research using cyberinfrastructure. Journal of Hydrometeorology, 16(3), 1155–1170.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. Educational Researcher, 32(1), 5–8.

Entekhabi, D., Njoku, E. G., O'neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., & Johnson, J. (2010). The soil moisture active passive (SMAP) mission. Proceedings of the IEEE, 98(5), 704–716.

Gochis, D., Yu, W., & Yates, D. (2015). The WRF-Hydro model technical description and user's guide, version 3.0. NCAR Technical Document, 120.

Haw, W., & Crawford, A. (2025). Using Hackathons to Enhance University-Level Water Curriculum for Students in Minority Communities. Journal of Learning Development in Higher Education.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., & Schepers, D. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K., & Iguchi, T. (2014). The global precipitation measurement mission. Bulletin of the American Meteorological Society, 95(5), 701–722.

Huppenkothen, D., Arendt, A., Hogg, D. W., Ram, K., VanderPlas, J., & Rokem, A. (2017). Hack Weeks as a model for data science education and collaboration. arXiv. https://arxiv.org/abs/1711.00028

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., & Woollen, J. (2018). The NCEP/NCAR 40-year reanalysis project. In Renewable energy (p. Vol1_146-Vol1_194). Routledge.

Kirkpatrick, D., & Kirkpatrick, J. (2006). Evaluating training programs: The four levels. Berrett-Koehler Publishers.

Krajewski, W. F., Ghimire, G. R., Demir, I., & Mantilla, R. (2021). Real-time streamflow forecasting: AI vs. Hydrologic insights. Journal of Hydrology X, 13, 100110.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., & Simpson, J. (1998). The tropical rainfall measuring mission (TRMM) sensor package. Journal of Atmospheric and Oceanic Technology, 15(3), 809–817.

Lara, M., & Lockwood, K. (2016). Hackathons as community-based learning: A case study. TechTrends, 60(5), 486–495.

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research: Atmospheres, 99(D7), 14415–14428.

Lighthill, M. J., & Whitham, G. B. (1955). On kinematic waves I. Flood movement in long rivers. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 229(1178), 281–316.

Likert, R., 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Nash, J. (1957). The form of the instantaneous unit hydrograph. Comptes Rendus et Rapports Assemblee Generale de Toronto, 3, 114–121.

Seibert, J. and Bergström, S., 2022. A retrospective on hydrological catchment modelling based on half a century with the HBV model. *Hydrology and Earth System Sciences*, *26*(5), pp.1371-1388.

Ramirez, C. E., Sermet, Y., & Demir, I. (2024). HydroCompute: An open-source web-based computational library for hydrology and environmental sciences. Environmental Modelling & Software, 175, 106005.

Ramirez, C. E., Sermet, Y., Molkenthin, F., & Demir, I. (2022). HydroLang: An open-source web-based programming framework for hydrological sciences. Environmental Modelling & Software, 157, 105525.

Richardson, L. F. (1922). Weather prediction by numerical process. Franklin Classics.

https://doi.org/10.5194/egusphere-2026-491
Preprint. Discussion started: 23 March 2026

855 Sherman, L. K. (1932). The relation of hydrographs of runoff to size and character of drainage-basins.
856 Eos, Transactions American Geophysical Union, 13(1), 332–339.
857 Seo, B. C., Keem, M., Hammond, R., Demir, I., & Krajewski, W. F. (2019). A pilot infrastructure for
858 searching rainfall metadata and generating rainfall product using the big data of NEXRAD.
859 Environmental modelling & software, 117, 69-75.
860 Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F., & Watkins, M. M. (2004). GRACE
861 measurements of mass variability in the Earth system. Science, 305(5683), 503–505.
862 Wagener, T., Savic, D., Butler, D., Ahmadian, R., Arnot, T., Dawes, J., Djordjevic, S., Falconer, R.,
863 Farmani, R., Ford, D., Hofman, J., Kapelan, Z., Pan, S., and Woods, R.: Hydroinformatics education – the
864 Water Informatics in Science and Engineering (WISE) Centre for Doctoral Training, Hydrol. Earth Syst.
865 Sci., 25, 2721–2738, https://doi.org/10.5194/hess-25-2721-2021, 2021.
866