TECHNICAL REPORT 22

July 2026

CUAHSI

allied for water science

www.cuahsi.org

THE UNIVERSITY OF ALABAMA® | Alabama Water Institute

Water Prediction Innovators Summer Institute Report 2026

Water Prediction Innovators Summer Institute Report 2026

Editors:

Nana Oye Djan

Megan Vardaman

Prepared in cooperation with the Consortium of Universities for the Advancement of Hydrologic Science, Inc.

CUAHSI Technical Report No. 22

Version 1.0

July 2026

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the opinions of NOAA.

Suggested Citation:

Vardaman, M., Djan, N., et al. (2026). Water Prediction Innovators Summer Institute Report 2026, HydroShare, http://www.hydroshare.org/resource/7b2b22d88a914bb79b3f48e5ea3ecaaf

Contents

Glossary of Terms

Term	Definition
AI	Artificial Intelligence
AWI	Alabama Water Institute
CIROH	Cooperative Institute for Research to Operations in Hydrology
CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science Inc.
FEMA	Federal Emergency Management Agency
FIM	Flood Inundation Map/Flood Inundation Mapping
HAND	Height Above Nearest Drainage
HAND-FIM	Height Above Nearest Drainage-Flood Inundation Map
ML	Machine Learning
MRMS	Multi-Radar Multi-Sensor
NextGen	Next Generation Water Resources Modeling Framework
NOAA	National Oceanic and Atmospheric Administration
NWC	National Water Center
OWP	Office of Water Prediction
QPE	Quantitative Precipitation Estimation
SRC	Synthetic Rating Curve
USGS	U.S. Geological Survey
WPI-SI	Water Prediction Innovators Summer Institute

Preface

This report summarizes the research conducted during the 2026 Water Prediction Innovators Summer Institute (WPI-SI). The WPI-SI is the result of a partnership between the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) and The University of Alabama. CUAHSI is a 501(c)(3) nonprofit organization with a mission "to empower the water community and advance science through collaboration, infrastructure, and education." The Cooperative Institute for Research to Operations in Hydrology (CIROH) funds this programming and supports the WPI-SI as a part of its mission to "to innovate research that enhances our understanding of how coupled atmosphere-oceanland-biosphere components interact hydrologically and transform this new knowledge into operational research products benefitting society." Funding for the Summer Institute is provided by the National Oceanic and Atmospheric Administration (NOAA), awarded to CIROH through the NOAA Cooperative Agreement with The University of Alabama, NA22NWS4320003.

Since 2015, The University of Alabama campus has hosted the WPI-SI. Graduate students come from different universities across the United States and work alongside academic faculty, federal scientists, researchers, private companies, and contractors in Tuscaloosa over the seven-week program. The research themes that students work under are ideated with input from the Office of Water Prediction (OWP) which has been supportive of this program from its inception.

Student fellows formed teams based on shared research interests and worked under the guidance of theme leaders throughout the program. This year, seven teams were established, each focused on a specific research question. Projects included characterizing the effect of the stage-dependent variability of Manning's roughness on Height Above Nearest Drainage-Flood Inundation Maps (HAND-FIM), reducing locational and seasonal biases in National Water Model (NWM) discharge estimates using machine learning algorithms, and establishing a pipeline for quantifying the sensitivity of actionable flash flood predictions to precipitation forcing uncertainty and spatiotemporal resolution.

In addition to their research, participants took part in social events and outings, forming lasting connections with their peers. Course Coordinators, both of whom were fellows in the past, offered assistance and support throughout the summer. For the fellows, the Summer Institute is more than a research program. It fosters teamwork, networking, and the development of lasting friendships.

Fellows

The 2026 cohort included 24 graduate students from 21 different universities across the United States. They began their journey with virtual meetings, which culminated in a two-week bootcamp where they participated in training on coding, data retrieval, and key topics related to their research projects. The research teams, formed at the end of the bootcamp, were mentored by theme leaders and worked to complete their research projects while giving weekly presentations to share their progress and receive feedback and guidance. The program culminates with this report and a Capstone Event wherein research teams present their work to a broad audience of researchers and peers. The fellows and theme leaders hail from various academic departments, including civil engineering, computer science, public policy, and earth sciences.

Themes and Theme Leads

The WPI-SI 2026 themes and theme leads were:

- The "Advancing the Use of Quantitative Precipitation Inputs into the NextGen Framework" theme, led by Mohamed Abdelkader (University of Iowa) and Humberto Vergara (University of Iowa).
- The "Innovations in Flood Inundation Mapping for Operational Forecasting" theme, led by Anupal Baruah (The University of Alabama) and Sagy Cohen (The University of Alabama).
- The "Advancing Machine Learning and Artificial Intelligence Frameworks for Enhanced Hydrologic Prediction" theme, led by Mohmmad Ali Javidian (Appalachian State University) and Sushant Mehan (South Dakota State University).
- The "Advancing Risk Communication and Public Engagement in Hydrologic Hazard Forecasting" theme, led by Sagy Cohen, (The University of Alabama) and Kelsey McDonough (The Water Institute).

Project Summaries

The 2026 WPI-SI projects are summarized below. Chapters 1-7 present the complete reports.

1. Projects within the "Advancing the Use of Quantitative Precipitation Inputs into the NextGen Framework" Theme:

Chapter 1: "Differentiable Modeling Leverages Sub-Hourly Radar Input for High-Performance Flash Flood Forecasting in NextGen". This report augmented the Next Generation Water Resources Modeling Framework (NextGen) suite of models through application of a sub-hourly differentiable rainfall-runoff model for flash flood prediction. The findings established sub-hourly differentiable modeling with Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation (QPE) as a working baseline for flash flood simulation with the NextGen framework.

Chapter 2: "Ensemble Streamflow Forecasting for Real-Time Flash Flood Prediction within the NextGen Framework". This report produced ensemble streamflow forecasts from a single deterministic simulation within the NextGen framework to test QPF location uncertainty in flash flood streamflow forecasts using the Neighboring Catchment Ensemble Technique (NCET), an extension of the Neighboring Pixel Ensemble Technique (NPET). The results showed that storm location, rather than donor similarity, controls the ensemble, offering a computationally efficient method for flash flood forecasting within the NextGen framework.

Chapter 3: "Sensitivity of Flash Flood Predictions to Precipitation Forcing Uncertainty for Different Spatiotemporal Resolutions". This report established a pipeline for quantifying the sensitivity of actionable flash flood predictions to precipitation forcing uncertainty and spatiotemporal resolution. Flash flood case studies were simulated in NextGen using sub-hourly MRMS precipitation. Uncertainty in peak flow magnitude bias and timing lag were evaluated by iteratively modifying precipitation magnitude scaling, spatial resolution, and temporal resolution. The results showed that the sensitivity of flash flood magnitude and timing predictions to precipitation forcing uncertainty responds only modestly to improvements in spatiotemporal resolution beyond some critical threshold.

2. "Projects within the Innovation in Flood Inundation Mapping for Operational Forecasting" Theme:

Chapter 4: "Evaluating the Sensitivity of HAND Flood Inundation Predictions to River Slope Dynamics". This report compared flood maps generated using existing Height Above Nearest Drainage-Flood Inundation Map (HAND-FIM) static river slopes and three variants of dynamic river slopes from SWOT (i.e., median, 90th percentile, maximum water surface elevation). Simulations were performed for six areas in the United States using NWM streamflow data. The results demonstrated the feasibility of dynamic, discharge-dependent slope estimation for future operational flood inundation mapping frameworks.

Chapter 5: "Improving Flood Inundation Mapping through Stage-Dependent Manning's n". This report determined whether a spatially and stage-dependent representation of Manning's n improves the accuracy of the operational HAND-FIM framework, relative to the fixed two-level (channel-floodplain) split or the single stage-invariant correction factor currently used in the National Water Model (NWM). The results showed that by correcting the channel-bed datum and calibrating roughness across five recurrence-based stage zones, median discharge error at held-out rating-curve points fell from 74.5% to 3.6%. While gauge-based calibration substantially improved rating-curve accuracy, its benefit for predicting FIM was inconsistent.

3. "Project within the Advancing Machine Learning and Artificial Intelligence Frameworks for Enhanced Hydrologic Prediction" Theme:

Chapter 6: "Advancing ML & AI Frameworks for Enhanced Hydrologic Prediction". This report tested whether different machine learning (ML) algorithms (i.e., simple recurrent network, gated recurrent unit, long short-term memory network, Transformer) can reduce the systematic local bias of the NWM and identified where these corrections had the greatest influence on streamflows. Models were trained under two setups: a residual setup that learned the difference between NWM discharge and observed discharge and added the learned correction back to NWM, and a direct setup that predicted observed discharge directly rather than a correction to NWM. Results from three gauged basins found that ML correction added the most value where NWM bias was highest.

4. "Project within the Advancing Risk Communication and Public Engagement in Hydrologic Hazard Forecasting" Theme:

Chapter 7: "Toward Better Flood Risk Communication: Applying a Diagnostic Framework for Evaluating and Redesigning FIM Visualizations". This report developed and applied a Diagnosing, Assessing, Redesigning, Testing, and Synthesizing (DARTS) framework to evaluate four FIM platforms for visualization clarity, interpretation burden, impact communication, and actionability. Findings were translated into design targets for a National Weather Prediction Service (NWPS) Emulator. The most persistent barriers identified were unclear terminology, weak first-screen clarity, layer and legend confusion, forecast timing and limitation gaps, and uneven impact translation.

Acknowledgments

Since the first Summer Institute in 2015, more than 250 graduate student fellows have collaborated on research projects focused on advancing the water prediction capabilities for the United States. Now in its 11th year, the Summer Institute has grown into a nationally recognized program that brings together emerging scientists, engineers, and professionals to develop innovative solutions for water-

related challenges. The 2026 WPI-SI builds on this legacy and would not have been possible without the commitment and hard work of many people:

We would like to express our sincere gratitude to the CUAHSI staff whose efforts were instrumental in making this program a reality: Jordan Read, Julia Masterman, Emily Clark, Summer Conley, Irene Garousi-Nejad, Tony Castronova, and Danielle Tijerina-Kreuzer. We give special thanks to the Alabama Water Institute, CIROH, and the Department of Geography and the Environment at the University of Alabama for hosting and supporting us this summer, especially Sagy Cohen, Lanna Nations, and Hannah Holcomb for handling logistics, including travel, housing, and campus coordination. Special appreciation goes to the Theme Leads who proposed the project ideas, mentored the teams, led the bootcamp sessions, and worked and supported us throughout the summer (and likely beyond).

Technical training was generously provided by Irene Garousi-Nejad, Tony Castronova, and Danielle Tijerina-Kreuzer (CUAHSI); Arpita Patel and Quinn Lee (CIROH); Taher Chegini (Dewberry); Whitney Flynn and Kevin Kalbaugh (FEMA); Daniel Rydl (Florida Division of Emergency Management); Nathan Young (Iowa Flood Center); Ryan Spies and Arash Rad (Lynker); Race Clark (NOAA National Severe Storms Laboratory); and Supath Dhital, Dipsikha Devi, Dan Tian, and Yixian Chen (University of Alabama); Melissa Kenney and Sajani Kandel (University of Minnesota). Thank you to the SI alumni, Danielle Tijerina-Kreuzer (CUAHSI), Irene Garousi-Nejad (CUAHSI), Mohamed Abdelkader (University of Iowa), and Whitney Flynn (FEMA), for sharing their experience and advice. CIROH and the Alabama Water Institute (AWI) offered multiple training sessions and were crucial in addressing numerous technical questions throughout the seven weeks of the Summer Institute. We thank the CIROH Science and Technology Team (especially Arpita Patel) for their guidance and for providing technical assistance and support with computational resources.

We also thank Ed Clark, Director of the National Water Center (NWC) and Deputy Director of the OWP and Steve Burian, Executive Director of CIROH and Director of Science at the Alabama Water Institute, for sharing their valuable insights and extensive experience with the fellows. Thanks to Fred Ogden, Carson Pruitt, and Jason Regina from the NWC for their technical support and engagement throughout the program.

We extend our sincere gratitude to everyone whose name may not appear here, but who contributed to this program's great success.

Thank you to everyone who made a contribution to the WPI-SI.

Nana Oye Djan

Student Course Coordinator, Water Prediction Innovators Summer Institute 2026

Ph.D. Student, Carnegie Mellon University

Megan Vardaman, Ph.D.

Student Course Coordinator, Water Prediction Innovators Summer Institute 2026

University of New Hampshire

Chapter 1:
Differentiable Modeling Leverages Sub-Hourly Radar Input for High-Performance Flash Flood Forecasting in NextGen

Jessica Keiser[1], Leo Lonzarich[2], Azizur Rahman[3], Mohamed Abdelkader[4*], and Humberto Vergara[5*]

[1]San Diego State University*; jkeiser6289@sdsu.edu*
[2]The Pennsylvania State University*; lgl5139@psu.edu*
[3]University of Texas at Arlington*; azizur.rahman@uta.edu*
[4]University of Iowa*; mohamed-abdelkader@uiowa.edu*
[5]University of Iowa*; humberto-vergaraarrieta@uiowa.edu*
*Theme Leader

Abstract: Rainfall-driven flash floods develop over scales of minutes to hours, yet many models, including those being considered for inclusion in the National Oceanic and Atmospheric Administration Office of Water Prediction (NOAA-OWP) Next-Generation National Water Model (NWM), run at hourly timesteps too coarse to capture the flashy precipitation and runoff responses driving these events. This work demonstrates the need for high-resolution precipitation inputs through the development and application of a sub-hourly differentiable rainfall-runoff model, using the bucket-based Hydrologiska Byråns Vattenbalans (HBV) model as a physical backbone. To facilitate this assessment, we assemble a continental-scale catalogue of nearly 20,000 flash flood events/reports from the NOAA Storm Events Database and leverage United States Geological Survey (USGS) gage observations to identify unreported events, all based on the NOAA Next Generation Water Resources Modeling Framework (NextGen) Hydrofabric v2.2 catchment network. For each event, we compile 2-minute precipitation data from the Multi-Radar Multi-Sensor (MRMS) system, hourly meteorological forcings from the Analysis of Record for Calibration (AORC) dataset, and USGS gage discharge observations to assemble a high-fidelity flash flood product for benchmarking and calibrating models on extreme events. We therefore develop a zonal-statistics workflow to translate 2-minute MRMS Quantitative Precipitation Estimation (QPE) from its native grid catchments in the NextGen Hydrofabric. We further develop a robust evaluation pipeline to assess the capacity for models to capture event peaks, total flow volume, and time-to-peak, among other metrics. Across the 1,132 flash flood events at Upper Neuse gages, dHBV2-Flash captured event peaks with a median peak flow bias of 16% and a median volume bias of 11.5%, with remarkably better performance skill at natural sites than at regulated ones. These findings establish sub-hourly differentiable modeling with MRMS QPE as a working baseline for flash flood simulation with the NextGen framework.

1. Motivation

Flash floods are among the deadliest weather hazards in the United States [1]. By definition, a flash flood rises and falls within hours of intense rainfall over a small area, with little to no advanced warning [2]. The July 4, 2025, Texas Hill Country disaster made the stakes plain: 119 people died in Kerr County, the deadliest flash flood in the United States since 1976 [3]. Flash flood prediction remains

difficult because the discharge response to precipitation in small basins develops within minutes to a few hours, and because small errors in rainfall timing or intensity distort the simulated flood disproportionately.

Time resolution is central to this problem. A model forced at an hourly timestep misses the short, sharp rainfall pulses that build flash flood peaks in small basins. The limitation is scale dependent: rainfall input skill that is adequate for large rivers loses most of its value in small basins where flash floods form [4]. NOAA's Office of Water Prediction (OWP) is transitioning from the National Water Model (NWMv3.0) to NextGen, a model-agnostic, standards-based system for deploying hydrologic formulations [5]. Current NextGen rainfall-runoff models run on hourly inputs, often too coarse for the timescales at which flash floods occur.

Operational systems already monitor flash floods at the scale at which they occur. Flash Flood Guidance compares observed and forecasted rainfall against threshold depths that would initiate flooding [6]. Flooded Locations And Simulated Hydrographs (FLASH) forces a distributed hydrologic model with MRMS rainfall to produce flash flood products at 1 km resolution every 2 minutes [7], and it correctly identified the location and magnitude of the Texas event [3]. Both systems are built for detection and situational awareness. Neither was designed to deliver discharge hydrographs along a continuous channel network, and hydrographs are what downstream services consume. Flood inundation mapping (FIM), which the National Weather Service (NWS) is deploying nationally, translates channel discharge into maps of inundation extent [8]. NextGen provides exactly that backbone: its Hydrofabric resolves the Contiguous United States (CONUS) into more than 800,000 catchments on a connected channel network [5], so every reach can carry a simulated hydrograph that FIM can ingest. What NextGen lacks is the sub-hourly capability that rapid onset flooding demands. This study is about augmenting NextGen with that capability, and with it, extending the reach of FIM toward the flash flood scale.

Three advances make the augmentation practical. MRMS dual-polarization algorithms deliver radar rainfall on a 1 km grid every 2 minutes [9]. Differentiable hydrologic models learn parameters from discharge observations by gradient descent while preserving process interpretability. Compiled languages such as Rust keep sub-hourly simulation affordable at operational scale. A sub-hourly differentiable model was built for the NextGen framework.

2. Objectives and Scope

The main goal of this study is to augment the NextGen suite of models with robust, sub-hourly capabilities for flash flood prediction in support of early warning systems. As proof-of-concept work, study scope is restricted to a selection of HUC8 drainage areas in CONUS, using water years 2021 to 2025, and gaged basins smaller than 1000 km^2 (the upper limit for flash flood scale basins [10]). The specific objectives are:

(1) To develop a high-fidelity CONUS-scale benchmark dataset tailored for flash flood modeling experiments, including model calibration and evaluation.
(2) To prototype a sub-hourly, state-of-the-art differentiable rainfall-runoff model within NextGen for flash flood simulations.
(3) To test the computational feasibility of sub-hourly simulation in operational settings.
(4) To develop an event-based evaluation framework to quantify flash flood forecast skill.

3. Previous Studies

NextGen replaces the monolithic WRF-Hydro architecture of the current NWM with a model-agnostic framework in which heterogeneous formulations run side by side on the CONUS Hydrofabric through a Basic Model Interface (BMI) [5], [11]. Its calibration practice, however, is built around the AORC dataset, an hourly meteorological forcing designed for long-term calibration and retrospective simulations [12]. The rainfall pulses that build flash flood peaks in small basins are aggregated away before any model on the framework sees them.

Radar-driven flash flood prediction at the national scale is not new. Since 2016, FLASH has forced the Coupled Routing and Excess Storage (CREST) distributed model [12] with MRMS precipitation [9] to map flash flooding across CONUS in real time [13], and it remains the operational reference for detection at flash flood scale. Two design choices limit its use for what we attempt here. Its parameters are set a priori rather than learned from discharge observations. Its products live on their own 1 km grid, so their simulated flows cannot propagate into high resolution channels such as those used for FIM services.

Differentiable hydrologic modeling addresses the first of those limits: process-based structure with neural network parameterization, trained end to end on discharge observations [14] [12]. Regionalized differentiable HBV matches Long Short-Term Memory (LSTM) Neural Network (NN) benchmarks while retaining interpretable states and parameters [15], [16]. dHBV2 extends the approach CONUS-wide at roughly 37 km² resolution with explicit river routing [17], and recent variants improve skill on unseen extreme events [18]. Every implementation to date, however, runs at daily or hourly steps outside operational frameworks. The large-sample datasets that enabled this progress, such as the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS), are likewise daily or hourly, and basin-mean [19]. No benchmark pairs sub-hourly radar rainfall with gauge discharge at event scale on an operational network. In this study, Radar and USGS Networked Observations for Flash Flooding (RUNOFF) and dHBV2-Flash are built to fill these two gaps.

4. Methodology

4.1. Flash-flood event catalog

We built the catalog from NOAA Storm Events Database records for 2021 to 2025 [20]. The period was set to overlap the dual-polarization algorithm in the MRMS QPE product, so every event pairs with radar precipitation of consistent quality. Database records often describe the same storm, so we aggregated them to the event level on two criteria. First, each record centroid is buffered using its start and end locations, and overlapping buffers merge. Second, records sharing an Episode ID, the database identifier linking reports from one meteorological system, are grouped. This yields 19,993 unique flash-flood events across CONUS and locates the regions most vulnerable to flash floods and to repetitive events.

We kept only events with an upstream or downstream USGS gauge, so each simulated hydrograph has a reference observation. Gages come from the Geospatial Attributes of Gages for Evaluating Streamflow (GAGES-II), provided by the US Geological Survey [21], restricted to flash flood scale of drainage areas under 1000 km². Event centroids and gauges were joined to the NextGen Hydrofabric. Gages were snapped to the nearest flowpath, each event assigned to the catchment holding its centroid, and that flowpath traced upstream and downstream to the nearest gauge. This reduced the catalog to 7,990 events. We then filtered on MRMS radar coverage. Coverage is a proxy for QPE

uncertainty [22]. It reflects how well the radar sees the precipitation that reaches the ground, not whether radar data exist. Events were retained when at least 80% of the upstream catchment was covered, which keeps events whose precipitation forcing is of adequate quality [23]. The final catalog holds 5,076 flash-flood events.

4.2. Study area and data

We selected USGS Hydrologic Unit Code 8 (HUC8) watersheds as the spatial unit for this study because they represent a standardized, nationally recognized watershed classification. The flash flood event catalogue was spatially joined with the CONUS HUC8 boundaries to determine the number of events occurring within each watershed. Watersheds were then ranked by event frequency, and the 15 HUC8 basins containing the greatest number of flash flood events were selected as the primary study areas. The Upper Neuse Basin in North Carolina is the top-ranked basin with 99 events (**Figure 1**) and was subsequently used for the first pass of benchmark model runs discussed in the remainder of this report.

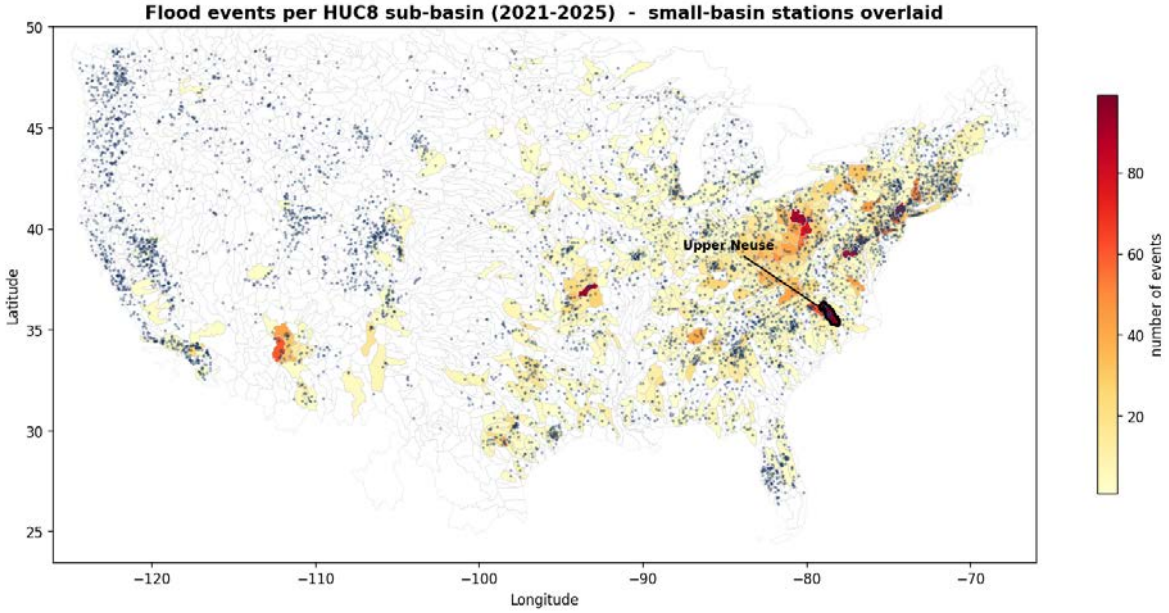


Figure 1: *Heatmap illustrating the distribution of events across CONUS HUC8 basins between WY2021-2025.*

Precipitation forcing comes from MRMS PrecipRate, the radar-only rain rate product, available in near real time at 2-minute, 0.01-degree resolution [9]. Because no existing routine mapped MRMS QPE onto the NextGen Hydrofabric, we developed a zonal-statistics geoprocessing workflow that aggregates gridded MRMS QPE into NextGen catchments, preserving precipitation volume and accounting for each catchment geometry (see Supplementary Materials).

4.3 dHBV2-Flash

To efficiently leverage sub-hourly precipitation forcing for flash flood and early warning modeling tasks at scale, we deploy a multi-timescale, distributed differentiable hydrologic model, dHBV2-Flash, inspired by the hourly-resolution multi-timescale introduced by Yang et al. [24]. For the demonstrative purposes of this study, we use the bucket-based Hydrologiska Byråns Vattenbalans (HBV) model as a basic conceptual backbone. HBV has 13 parameters, 3 of which are learned dynamically in time with an LSTM, and the remaining 10 of which are learned statically in time with a Multi-Layer Perceptron

(MLP). These NNs are then trained end-to-end through HBV using automatic differentiation in PyTorch [25] and with USGS gage observations as ground truth constraining learning. In the multi-timescale formulation, two differentiable models of this structure are coupled together: a "low-frequency" (hourly) model is used to warm up antecedent HBV and LSTM states with hourly forcing input; initialized states are then transferred to a "high-frequency" (sub-hourly) model, which delivers 15-minute streamflow simulations for a desired time window. Both models are distributed on HydroFabric catchments [17], allowing for high-resolution regionalization, and per-catchment discharge is routed to gaged outlets. Critically, this structure allows efficient scalability and throughput on both Graphical Processing Units (GPUs) and Central Processing Units (CPUs). Further details on model structure and developments can be found in Supplementary Materials.

4.4. Model Evaluation

The model evaluation pipeline utilizes USGS discharge data from gages inside the study catchments as ground truth. The temporal resolution of the discharge data is aggregated to match the 15-minute temporal resolution of the dHBV2-Flash model outputs. The design of the evaluation pipeline is based on the NOAA-OWP HydroTools event separation package [26] and event matching methodology [27], [28] to quantify how accurately the model is simulating the observed events, namely peak magnitude and timing. Events are detected independently in both the observed and simulated hydrographs using the HydroTools decomposition algorithm. Observed events whose peak discharge equals or exceeds the Q50 (flow magnitude exceeded 50% of the time) of the full observed flow duration curve are retained for evaluation, providing a consistent selection threshold across catchments. Matched pairs are identified using an interval matching algorithm, which places a 12-hour gate around the beginning and end of each observed event peak. The algorithm selects the best simulated candidate via a cost function that weights timing proximity and log-magnitude difference. Per-event metrics computed for each matched pair include peak discharge error, peak timing error, and runoff volume error, with site-level summaries reporting the median and interquartile range across all detected events.

Model performance is evaluated on median peak discharge bias, median volume bias, and median timing bias. Performance metrics including the Richards-Baker Flashiness Index difference (RBI diff) [29] and Nash-Sutcliffe efficiency (NSE) [30] are also calculated. Evaluations were performed on the newly developed model and model simulations obtained from uncalibrated CFE models forced by hourly AORC and MRMS QPE products to investigate the impact of different QPE sources on flow simulation in flash flood conditions. We refer readers to Supplementary Materials for a detailed description of the experiment.

5. Results

5.1. RUNOFF v1.0

Our work on the flash flood event catalogue (section 4.1) has led to the development of RUNOFF v1.0 (Radar & USGS Networked Observations of Flash Floods). Building on the NOAA Storm Events Database first developed at the national scale by Gourley et al. [31], RUNOFF v1.0 extends this foundation to be directly applicable to flash flood modeling experiments. The dataset is built on the NextGen HydroFabric v2.0 and can be aggregated to the USGS HUC8 level, improving operational compatibility in multiple formats. This addresses the current gap in large-sample, simulation-compatible flash flood datasets. With 5,097 unique flash flood events across CONUS, each paired with quality-controlled MRMS QPE forcing and a reference USGS gage, the dataset is designed

to work for model calibration, evaluation, and validation. Access to the dataset is available in the Supplementary Materials.

5.2. dHBV2-Flash Performance

dHBV2-flash was evaluated for 1,132 chosen simulated events from the end of 2024 to 2025 across all 12 gage sites in the Upper Neuse subbasin. The median peak flow bias across sites was +16%, while the median peak timing error ranged from -8 to 0 hours, indicating a tendency for the model to overpredict peaks at or before the observed time. The median volume bias was +11.5%, meaning volume was also generally overpredicted. The Richards-Baker Flashiness Index difference (RBI diff) is used in the evaluation to measure how accurately a model is simulating observed response to rainfall. A positive value for this metric means that the simulation is flashier than observed. With most sites reporting values near zero, the model captures flashiness in the hydrograph reasonably well. Nash-Sutcliffe efficiency (NSE) was used to measure overall predictive skill in the model, with a value less than zero indicating the model performed no better than using the observed mean. The NSE values were negative at all sites, indicating the simulations did not outperform the mean observed flow.

When three dam-regulated sites (02085000, 02086500, 02086624) were excluded, median peak flow and volume bias were largely unchanged, at +15.5% and +11.1%, respectively. RBI diff values also remained very close to zero. However, the median NSE increased significantly, from -2.12 to -0.45. This suggests that dam sites were responsible for the most inaccurate simulated hydrograph shapes. Timing accuracy also increased slightly, with the worst median timing error decreasing from -8 to -6.75 hours. These results suggest that, while dam regulated sites contributed to poor predictive performance, other biases revealed in the evaluation (peak discharge, volume, and timing) may come from other sources.

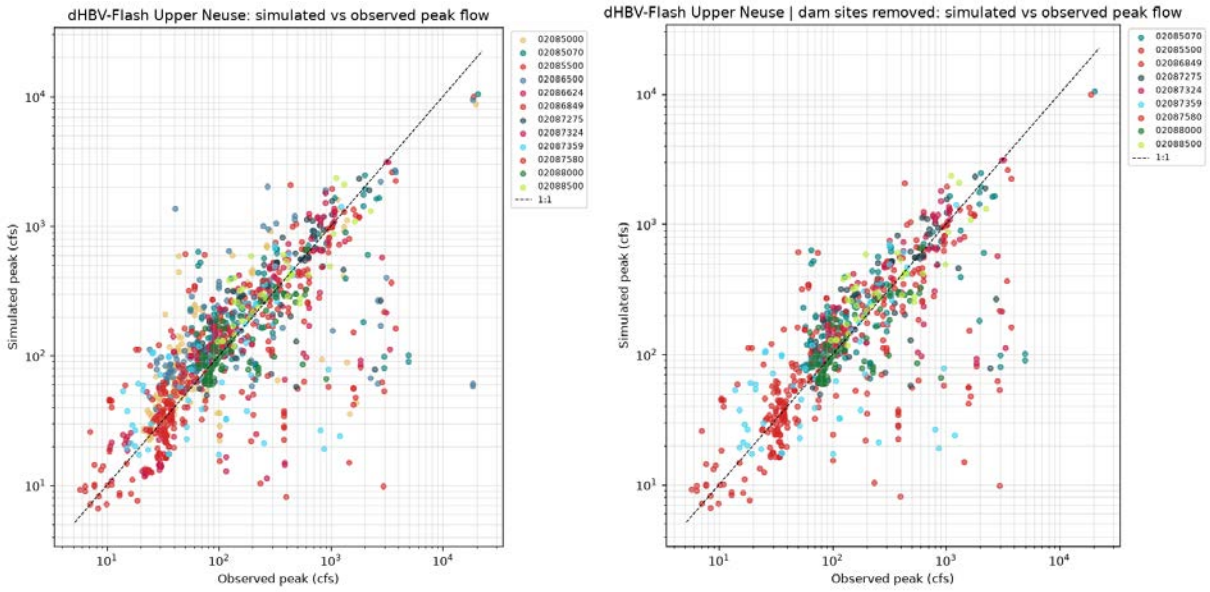


Figure 2: *Log-scale scatter plots showing the general performance of dHBV-Flash peak prediction.*

Figure 2 shows the distribution of simulated vs. observed peak flows for all events in Upper Neuse vs. with the dam sites removed. The plots illustrate the general positive peak flow bias across all sites. Visually, the plot suggests that performance was more reasonable at higher peak flow magnitudes, as scatter increases significantly in lower-flow events. Removing the dam sites does not significantly

affect performance for peak prediction, but does remove some major outliers. **Figure 3** shows the full distribution of bias values for peak magnitude and timing across all events in relation to drainage area. Median peak flow bias sits near zero for drainage areas of ≤ 300 km^2, meaning dHBV2-Flash is getting the magnitude right for most events, despite outliers on both sides. Peak flow bias drifts positive with increasing catchment size, which may reflect more compounded error due to more aggregated runoff from subcatchments. The peak flow bias also has a hard floor at -100%, which the whiskers touch, meaning that some observations were essentially completely missed by the simulation. Peak timing errors show the opposite trend, with smaller basins having significantly more bias towards early prediction. This timing degradation in the smallest catchments, which are most affected by flash floods, highlights why sub-hourly QPE forcing is important to improve model timing performance.

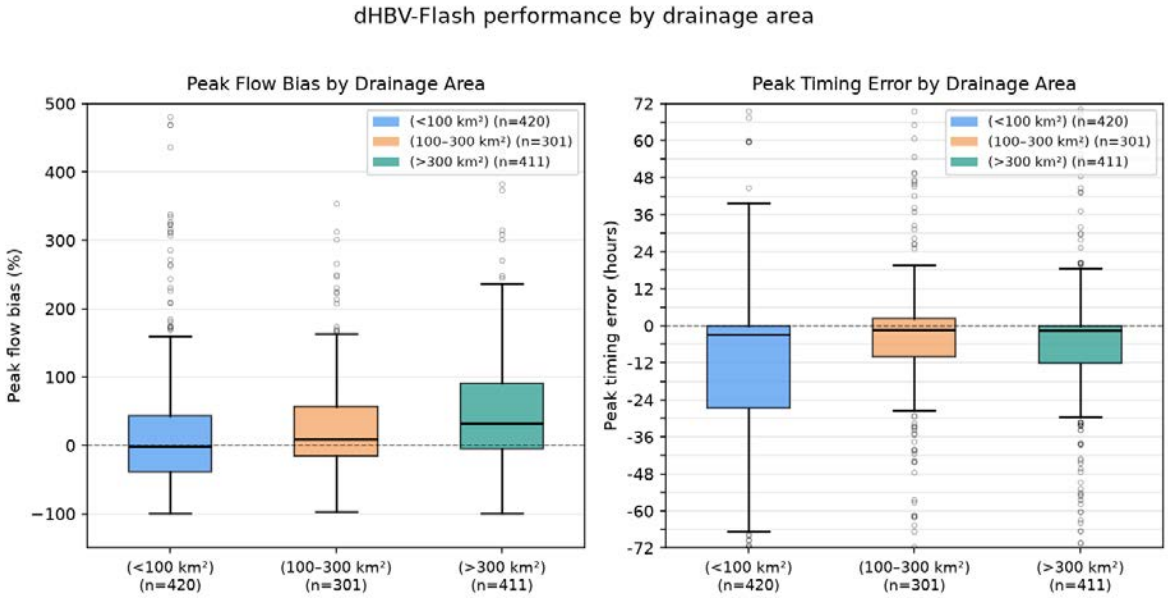


Figure 3: *Box-and-whisker plots showing peak flow bias and peak timing error in relation to drainage area.*

6. Conclusion

The flash floods of Kerr County, Texas, and many other high-profile events, are reminders that the standard hourly scale of national water modeling infrastructure is inadequate for resolving extreme events at the flash flood scale. This work intends to address this gap, and motivate future work, by establishing a baseline for operational flash flood modeling development and evaluation.

To address the lack of established flash flood-specific datasets, we introduced RUNOFF v1.0, a scalable, continental-scale catalog of over 5,000 flash flood events as a community resource for flash flood model development. Built on the NextGen Hydrofabric for operational compatibility, it provides flash flood event identification from the NOAA Storm Events Database, associated USGS gages, and quality-controlled, high-resolution MRMS precipitation forcing. Following and providing this resource, we also show that sub-hourly MRMS forcing can be routinely and efficiently mapped onto NextGen HydroFabric catchments in both research and operational settings (i.e., within NextGen).

To leverage this input, we also implemented a sub-hourly, distributed, differentiable model, dHBV2-Flash, for which preliminary evaluations on the Upper Neuse HUC8 demonstrated the capability of

this approach for capturing flash flood peak magnitudes and timings. It is critical to note that developments of this model and expanded application to a continental scale are ongoing.

This work arrives at a point where sub-hourly flash flood modeling for future NOAA-OWP operations remains largely unaddressed: existing radar-driven systems such as FLASH were not built on the NextGen Hydrofabric, and existing NextGen-ready models have so far been developed exclusively for daily or hourly resolution. The concurrent maturation of dual-polarization MRMS, state-of-the-art differentiable modeling methods, and the NextGen framework itself makes this a timely opportunity to close this gap on infrastructure with high-impact potential, before operational architecture decisions are finalized.

Together, these results suggest that sub-hourly, differentiable modeling is not just feasible within NextGen, it is necessary if the framework is going to serve flash flood early warning at the scale FIM and downstream inundation products require. The preliminary work here suggests several priorities for future development: extended calibration and evaluation across all of CONUS to test regional transferability; investigate reservoir and dam-aware routing to correct the largest source of poor hydrograph recovery; and address systematic bias through smarter physical model backbones designed for flash floods at small scales, and which account for environmental factors like impervious surfaces and urban drainage.

Acknowledgements: We would like to thank Dr. Fred Ogden for his guidance with CFE; Quinn Lee and Josh Cunningham for supporting the setup of NOAA's NextGen architecture; Jason Regina for providing hydrograph separation methodology via HydroTools. We would also like to thank Dr. Chaopeng Shen and Dr. Yalan Song for their insights developing the modeling scheme proposed in this work. Additional thanks to Dr. Michelle Hummel and Dr. Hilary McMillan for their professional recommendations and continued support.

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA.

This research utilized CIROH Cyberinfrastructure external resources - NSF ACCESS allocation number EES240087 on Indiana University's Jetstream2 managed by CIROH Cyberinfrastructure. ACCESS is supported by NSF awards #2138259, #2138286, #2138307, #2137603, and #2138296. The authors appreciate support from the CIROH Cyberinfrastructure team. Learn more: https://hub.ciroh.org/docs/services/intro.

Generative AI tool Claude Sonnet 4.6 was used in this report to refine grammar and clarity throughout. All factual content and interpretations were verified by the authors. This tool was also used to assist with drafting the data collection and analysis code, as well as debug the model. All code was reviewed, tested, and validated by the authors before use in this study.

Supplementary Materials: All supplementary materials can be found in the Hydroshare: Supporting Materials for Sub-Hourly Runoff Modeling in NextGen | CUAHSI HydroShare.

RUNOFF v1.0 was generated with the Runoff GitHub package. NextGen-ready dHBV2-Flash code (training code will be made available at a later time) are available in the dhbv2-flash NextGen module.

References

[1] National Weather Service, "Summary of Natural Hazard Statistics for 2024 in the United States," National Oceanic and Atmospheric Administration, Natural Hazard Statistics Report, Mar. 2026. Accessed: Jul. 17, 2026. [Online]. Available: https://www.weather.gov/media/hazstat/sum24.pdf

[2] "Flash flood - Glossary of Meteorology." Accessed: Jul. 17, 2026. [Online]. Available: https://glossary.ametsoc.org/wiki/flash-flood/

[3] "WoFS-FLASH coupled forecasts for the July 2025 Texas Hill Country Flash Flood Disaster in: Bulletin of the American Meteorological Society - Ahead of print." Accessed: Jul. 17, 2026. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/aop/BAMS-D-25-0252.1/BAMS-D-25-0252.1.xml

[4] "Is This Rainfall Forecast Good or Bad? For Flood Forecasting, the Answer Is Scale Dependent in: Bulletin of the American Meteorological Society Volume 106 Issue 9 (2025)." Accessed: Jul. 17, 2026. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/106/9/BAMS-D-24-0166.1.xml

[5] "The NextGen Water Resources Modeling Framework: Community Innovation at the Intersection of Hydrologic, Data and Computer Sciences - Ogden - 2026 - JAWRA Journal of the American Water Resources Association - Wiley Online Library." Accessed: Jul. 17, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/1752-1688.70089

[6] R. A. Clark, J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, "CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products," *Weather Forecast.*, vol. 29, no. 2, pp. 377–392, Apr. 2014, doi: 10.1175/WAF-D-12-00124.1.

[7] "The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United States in: Bulletin of the American Meteorological Society Volume 98 Issue 2 (2017)." Accessed: Jul. 17, 2026. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/98/2/bams-d-15-00247.1.xml

[8] F. Aristizabal *et al.*, "Extending Height Above Nearest Drainage to Model Multiple Fluvial Sources in Flood Inundation Mapping Applications for the U.S. National Water Model." Accessed: Jul. 17, 2026. [Online]. Available: https://repository.library.noaa.gov

[9] J. Zhang *et al.*, "Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities," *Bull. Am. Meteorol. Soc.*, vol. 97, no. 4, pp. 621–638, Apr. 2016, doi: 10.1175/BAMS-D-14-00174.1.

[10] Z. L. Flamig, H. Vergara, and J. J. Gourley, "The Ensemble Framework For Flash Flood Forecasting (EF5) v1.2: description and case study," *Geosci. Model Dev.*, vol. 13, no. 10, pp. 4943–4958, Oct. 2020, doi: 10.5194/gmd-13-4943-2020.

[11] "WRF-Hydro® Modeling System | Research Applications Laboratory." Accessed: Jul. 17, 2026. [Online]. Available: https://ral.ucar.edu/solutions/products/wrf-hydro-modeling-system

[12] J. Wang *et al.*, "The coupled routing and excess storage (CREST) distributed hydrological model," *Hydrol. Sci. J.*, vol. 56, no. 1, pp. 84–98, Feb. 2011, doi: 10.1080/02626667.2010.543087.

[13] J. J. Gourley *et al.*, "The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United States," *Bull. Am. Meteorol. Soc.*, vol. 98, no. 2, pp. 361–372, Feb. 2017, doi: 10.1175/BAMS-D-15-00247.1.

[14] "Differentiable modelling to unify machine learning and physical models for geosciences | Nature Reviews Earth & Environment." Accessed: Jul. 17, 2026. [Online]. Available: https://www.nature.com/articles/s43017-023-00450-9

[15] "Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. - Google Search." Accessed: Jul. 17, 2026. [Online]. Available: https://www.google.com/search?q=Differentiable%2C+learnable%2C+regionalized+process-based+models+with+multiphysical+outputs+can+approach+state-of-the-art+hydrologic+prediction+accuracy.&rlz=1C1GCEA_enUS1168US1168&oq=Differentiable%2C+learnable%2C+regionalized+process-based+models+with+multiphysical+outputs+can+approach+state-of-the-art+hydrologic+prediction+accuracy.&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIHCAEQIRiPAtIBBzY5NmowajSoAgGwAgHxBRa7JliLIT6E&sourceid=chrome&source=chrome.ob&ie=UTF-8

[16] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets," *Hydrol. Earth Syst. Sci.*, vol. 23, no. 12, pp. 5089–5110, Dec. 2019, doi: 10.5194/hess-23-5089-2019.

[17] "High-Resolution National-Scale Water Modeling Is Enhanced by Multiscale Differentiable Physics-Informed Machine Learning - Song - 2025 - Water Resources Research - Wiley Online Library." Accessed: Jul. 17, 2026. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024WR038928

[18] "Physics-Informed, Differentiable Hydrologic Models for Capturing Unseen Extreme Events - Song - 2026 - Water Resources Research - Wiley Online Library." Accessed: Jul. 17, 2026. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025WR040414

[19] "HESS - The CAMELS data set: catchment attributes and meteorology for large-sample studies." Accessed: Jul. 17, 2026. [Online]. Available: https://hess.copernicus.org/articles/21/5293/2017/

[20] "Storm Events Database | National Centers for Environmental Information." Accessed: Jul. 16, 2026. [Online]. Available: https://www.ncei.noaa.gov/stormevents/

[21] Falcone, James A., "GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow." U.S. Geological Survey, 2023. doi: 10.5066/P96CPHOT.

[22] J. J. Gourley *et al.*, "The FLASH project: Improving the tools for flash flood monitoring and prediction across the United States," *Bull. Am. Meteorol. Soc.*, vol. 98, no. 2, pp. 361–372, Feb. 2017, doi: 10.1175/BAMS-D-15-00247.1.

[23] J. J. Gourley *et al.*, "The FLASH Project: Improving the Tools for Flash Flood Monitoring and Prediction across the United States," Feb. 2017, doi: 10.1175/BAMS-D-15-00247.1.

[24] W. Yang, H. Ji, L. Lonzarich, Y. Song, and C. Shen, "Diffusion-based probabilistic modeling for hourly streamflow prediction and assimilation," Oct. 09, 2025, *arXiv*: arXiv:2510.08488. doi: 10.48550/arXiv.2510.08488.

[25] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, Oct. 2017. Accessed: Mar. 21, 2022. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ

[26] J. Regina and A. Raney, "OWPHydroTools: Python-based Tools for Retrieving and Evaluating National Water Model Streamflow Simulations," presented at the AGU Fall

Meeting Abstracts, Dec. 2021, pp. H45O-1343. Accessed: Jul. 14, 2026. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2021AGUFM.H45O1343R

[27] U. Ehret and E. Zehe, "Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events," *Hydrol. Earth Syst. Sci.*, vol. 15, no. 3, pp. 877–896, Mar. 2011, doi: 10.5194/hess-15-877-2011.

[28] J. Ewen, "Hydrograph matching method for measuring model performance," *J. Hydrol.*, vol. 408, no. 1, pp. 178–187, Sep. 2011, doi: 10.1016/j.jhydrol.2011.07.038.

[29] D. B. Baker, R. P. Richards, T. T. Loftus, and J. W. Kramer, "A New Flashiness Index: Characteristics and Applications to Midwestern Rivers and Streams," *JAWRA J. Am. Water Resour. Assoc.*, vol. 40, no. 2, pp. 503–522, 2004, doi: 10.1111/j.1752-1688.2004.tb01046.x.

[30] R. H. McCuen, Z. Knight, and A. G. Cutter, "Evaluation of the Nash–Sutcliffe Efficiency Index," *J. Hydrol. Eng.*, vol. 11, no. 6, pp. 597–602, Nov. 2006, doi: 10.1061/(ASCE)1084-0699(2006)11:6(597).

[31] J. Gourley *et al.*, "A Unified Flash Flood Database across the US," *Bull. Am. Meteorol. Soc.*, p. 130125093450003, Jan. 2013, doi: 10.1175/BAMS-D-12-00198.

Chapter 2:
Ensemble Streamflow Forecasting for Real-Time Flash Flood Prediction within the NextGen Framework

Amirhossein Montazeri[1], Mohammad Mosavat[2], Niloufar Soheili[3], Aldo Andres Tapia Araya[4], Mohamed Abdelkader[5*], and Humberto Vergara[6*]

[1]Boise State University, Boise, ID, USA; *amirhosseinmonta@u.boisestate.edu*
[2]The University of Alabama, USA; *smosavat@crimson.ua.edu*
[3]City University of New York (City College), New York, USA; *nsoheili@ccny.cuny.edu*
[4]The University of Arizona; *aldotapia@arizona.edu*
[5]The University of Iowa; *mohamed-abdelkader@uiowa.edu*
[6]The University of Iowa; *humberto-vergaraarrieta@uiowa.edu*
*Theme Leader

Abstract: Flash flood forecasting depends on quantitative precipitation forecasts (QPF), whose location errors can shift the simulated flood to the wrong basin. Representing this uncertainty with QPF ensembles requires one model run per member, which real-time operations cannot afford. The goal of this study is to produce ensemble streamflow forecasts from a single deterministic simulation within the NextGen framework. To this end, we developed the Neighboring Catchment Ensemble Technique (NCET), which extends the Neighboring Pixel Ensemble Technique (NPET) to the hydrofabric catchments by borrowing routed hydrographs from neighboring donors, weighted by storm displacement and physiographic similarity. Applied to a real HRRR forecast case, NCET recovered a flood that the deterministic forecast missed: isotropic weighting improved the CRPS skill score by about 0.30, and aiming the weights along the storm displacement improved the skill to about 0.53 - 0.60. This indicates that storm location, rather than donor similarity, controls the ensemble, offering a computationally inexpensive path to flash flood forecasting within the NextGen framework.

1. Motivation

Flash floods are among the most rapid and deadly natural hazards [1], and they account for most flood related fatalities in the United States [2]. The interval between the rainfall and the peak streamflow is often shorter than the time needed to issue and act on a warning, so accurate and timely forecasts of the hydrologic response are critical. Despite advances in forecasting and warning systems, fatalities have not declined in recent decades. Between 1996 and 2017, flash floods killed 1,399 people across the contiguous United States, an average of about 64 deaths per year [3].

Operational flood forecasting drives a distributed hydrologic model with quantitative precipitation forecasts (QPF) to extend lead time beyond the time a catchment takes to respond to rainfall [4]. This matters most for flash floods, which occur in small catchments, a few hundred square kilometers or less, and respond to rainfall within about six hours. Short-range QPF carries large uncertainty, and storm-location errors in particular can severely affect flash flood prediction. These errors can be on the order of 100 km [5], [6], with variation across regions and seasons [7], [8]. A displacement of this size places the simulated response in the wrong basins, and in small flash flood catchments it can exceed the drainage area itself. The catchment truly at risk then goes undetected, its peak streamflow is underpredicted, and warnings arrive late or not at all.

The standard response, forcing the model with a QPF ensemble, is too costly for the large domains and short cycle times of operational flash flood systems [9]. A practical method must instead carry precipitation-input uncertainty from a single hydrologic simulation. This study translates NPET from the pixel grid to the catchments of the NextGen hydrofabric, a method we call the neighboring catchment ensemble technique (NCET). Each catchment is treated as a target node, and neighboring catchments within a search radius as potential donors. Their hydrographs come from a single deterministic NextGen simulation, weighted by their spatial displacement from the target and filtered by drainage-area similarity. Beyond this baseline, we introduce two data-driven refinements. The first reweighs donors by physiographic similarity, so that catchments likely to respond like the target contribute more. The second transforms the borrowed hydrographs with a learned surrogate model to better match the target's expected response.

2. Objectives and Scope

This work develops and evaluates a low-cost post-processing framework that represents QPF location uncertainty in flash flood streamflow forecasts within NextGen, at the small catchment scale, without additional hydrologic model runs. The central idea is to generalize the NPET from the regular pixel grid on which it was formulated to the irregular catchment polygons of the NextGen hydrofabric, so that a probabilistic streamflow forecast can be produced from a single deterministic NextGen simulation.

The specific objectives are:

- To develop and implement NCET, a generalization of NPET to the NextGen hydrofabric, enabling ensemble streamflow forecasts.
- To quantify the sensitivity of ensemble performance to storm displacement, through controlled experiments with affine-transposed observed storms of known displacement.
- To assess whether data-driven refinements, attribute-informed donor weighting and a surrogate rainfall-runoff model, improve ensemble skill beyond the spatial and drainage area criteria of NPET.
- To quantify the probabilistic skill that NCET adds over the deterministic simulations.

The experiments were conducted in two parts: first with observed radar rainfall (MRMS) displaced in space by known amounts, to isolate the effect of storm location under controlled conditions, and then with a real HRRR forecast, to test the framework under operational QPF error.

3. Previous Studies

Traditional flash flood forecasting relied on deterministic products driven by observed rainfall. In operational settings, Flash Flood Guidance (FFG) gives each River Forecast Center the 1-, 3-, and 6-h rainfall that would cause flooding on small streams, and forecasters flag a threat when observed rainfall approaches that value [10], [11]. The FLASH system takes a more direct path, forcing the CREST hydrologic model with Multi-Radar Multi-Sensor (MRMS) rainfall to simulate streamflow nationwide at 1-km, 10-min resolution [12]. Both rest on precipitation estimates (QPE), and since QPE describes rain that is already falling, these systems leave forecasters little lead time.

More lead time comes from forecasted rainfall (QPF), which anticipates precipitation before it is observed [4], [13]. But a deterministic forecast is inherently affected by errors from different sources, so an honest use of QPF is to represent the plausible outcomes with an ensemble. At the convection-

allowing scale that flash floods demand, the operational ensemble is the High-Resolution Ensemble Forecast (HREF), which aggregates several independently designed convection-allowing models [14]. The deterministic short-range forecast most used in operations is the High-Resolution Rapid Refresh (HRRR), a 3-km convection-allowing model that updates every hour [15].

HRRR carries errors in the pattern, magnitude, timing, and location of rainfall [16]. For flash floods the location error matters most, because the rain must fall in the correct small basin to produce a response in the right location. An HRRR forecast can shift the rainfall field by roughly 100 km even when the predicted amount and storm type are reasonable [17]. A fixed displacement matters more for a small basin than a large one [10], so in a flash flood catchment it can move the rainfall out of the target basin entirely. The basin truly at risk then goes undetected while a neighbor receives a false signal.

The established way to account for this uncertainty in streamflow is to force the hydrologic model with an ensemble of QPFs, propagating each member through the model to obtain a streamflow ensemble. This approach improves reliability, but every precipitation member needs its own hydrologic run. Flash floods develop too quickly to afford one full run per member, so the repeated simulations can consume the lead time the forecast was meant to provide. Efficient post-processing can represent QPF location uncertainty from a single hydrologic simulation, at a fraction of the cost. The neighboring pixel ensemble technique (NPET) does this by building a streamflow ensemble from one QPF-forced run, borrowing the simulated hydrograph from a nearby pixel of similar physiography, weighted by distance and filtered by drainage area [8]. NPET works on a pixel-based mesh. NextGen, the basis of the next generation National Water Model, instead represents the landscape as catchments connected by flowpaths [18].

4. Methodology

The methodology consists of four components: (1) study area and data preparation, (2) deterministic streamflow simulation with the NextGen framework, (3) adaptation of NPET to the catchment scale as the NCET, and (4) probabilistic evaluation of the resulting ensemble approaches (CRPS).

4.1. Study areas and data

The Study focuses on the flash flood events in Ellicott City of May 27, 2018, where the QPF storm-location error affected the forecast of the upstream catchments of the streams that cross the city. The study domain was defined using a 200-km buffer around the target catchments because the NCET method represents precipitation and location uncertainty through a radial neighborhood search. All hydrofabric catchments located within and touch this buffer regardless of their position within the river network were included in the analysis (**Figure 1**). The resulting domain contained 8,474 catchment divides, along with their associated flowpaths, nexus points (computational junctions), drainage areas, and physiographic attributes, extracted from Vector Processing Unit (VPU) 02 of the CONUS NextGen hydrofabric. MRMS quantitative precipitation estimates were used as the reference forcing, while HRRR quantitative precipitation forecasts represented forecast rainfall and its associated spatial uncertainty.

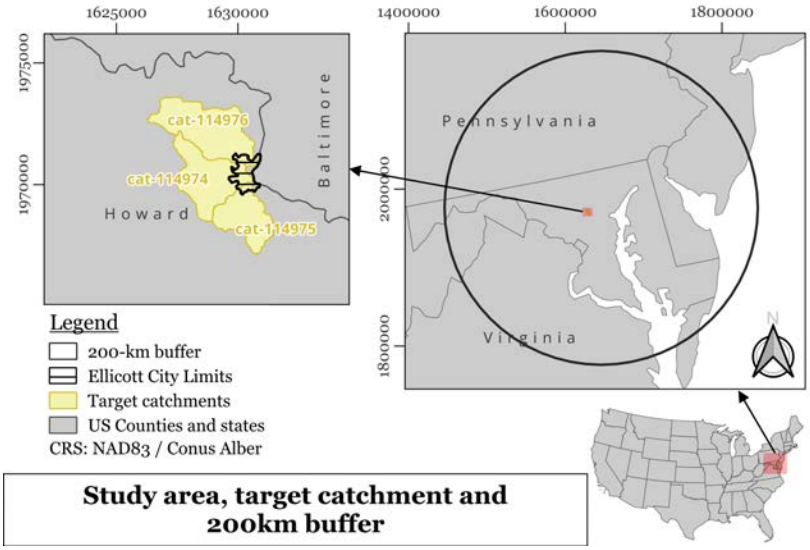


Figure 1. *Geographical location of Ellicott City, Maryland, the target catchments affecting the city, and the surrounding 200-km study-area buffer.*

4.2. Hydrologic modeling in NextGen and NGIAB

The NextGen framework [18] was used to: (1) download input forcings; (2) aggregate dynamic input variables; (2) hydrological modeling; (3) runoff routing. Since the NextGen data pre-processing library only have access to Analysis Of Record for Calibration (AORC) and National Water Model Retrospective datasets for forcings download [19], we modified the library to replace AORC precipitation with MRMS hourly precipitation and HRRR 18-hour forecast data. Since all these products have different spatial resolutions, a regridding with linear interpolation was applied to match the resolution using AORC data as the target resolution (1 arc-second). Then, the same framework was used to apply the zonal statistics geoprocessing to aggregate modified forcings data.

The NGIAB containerized Docker image was used to run the Conceptual Functional Equivalent (CFE) rainfall-runoff hydrological model [20], coupled with T-Route, a topological routing scheme, to obtain the streamflow of every catchment inside the study area. T-Route is designed to work with the hydrofabric database, routing the streamflow through the Nexus points available in the database and, in many cases, setting the outlet discharge of headwater catchments to 0 m3s-1 due to the absence of Nexus points in those catchments, limiting the hydrologic response sampling of headwater catchments under the current NextGen framework. To overcome this challenge, we extended NextGen using the NOAA Duamel Unit Hydrograph routing to compute the catchment discharge (Gamma Nash-Cascade convoluted by Duhamel integral). Details of the method are provided in the supplementary materials and in the OWP reference implementation (https://github.com/NOAA-OWP/sac-sma/blob/master/src/sac/duamel.f).

Finally, the modeled streamflow was used to evaluate the NCET method. The summary of the workflow, from data acquisition to obtaining the streamflow series to evaluate NCET, is presented in **Figure 2**. The workflow starts downloading the QPE (MRMS) and QPF (HRRR) data to replace the AORC precipitation data, followed by data aggregation. Since HRRR product length is a 18-h forecast at hourly time step, the data from this product replaces the time range of the MRMS product. This allows CFE to run a warm-up period using the observed precipitation product, and create two files: one with the modeled runoff using the QPE product; and, with the QPE followed by QPF, to

compute the differences between the forecasted precipitation and the observed by the Radar Network. NOAA Duamel routing scheme is used to obtain the modeled streamflow. Only the overlapping period of QPE and QPF is used to evaluate NCET.

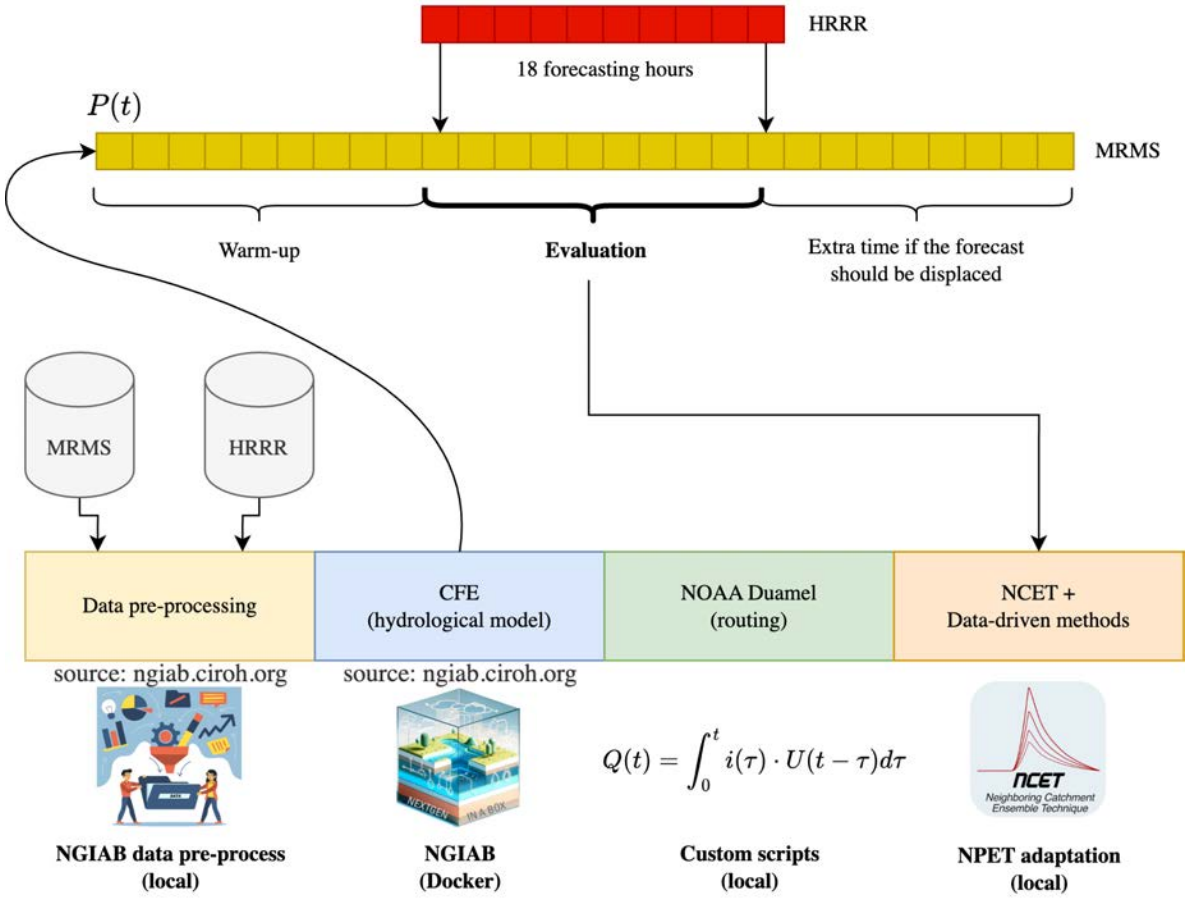


Figure 2. *Workflow summary for running and evaluating the NPET adaptation at catchment level. The source of the logos of NGIAB is ngiab.ciroh.org.*

4.3. Catchment-scale ensemble generation (NCET)

NCET translates NPET from a regular pixel grid to the irregular catchments of the NextGen hydrofabric. For each target catchment, potential donor catchments are identified within the defined search radius following the isotropic and anisotropic sampling scheme introduced in previous studies. In the isotropic scheme, catchments within the search domain are treated equally in terms of spatial displacement, and donor eligibility is determined primarily by drainage area similarity to the target catchment. In contrast, the anisotropic scheme represents some information about the direction of storm displacement using a bivariate beta weighting function, which assigns different spatial weights to donor catchments according to their locations relative to the target. The hydrographs of the retained donor catchments are then used to construct a probabilistic ensemble from a single deterministic NextGen simulation, **Figure 3** represent both schemas.

Although the original NPET framework captures QPFs uncertainty, there remains potential to improve ensemble accuracy by better informing the algorithm about which donor catchments are most likely to produce hydrologic responses similar to the target catchment. We therefore introduce

two data-driven refinement approaches. In both cases, a drainage-area filter retains only donor catchments with contributing areas sufficiently similar (Eq. (1)) to that of the target.

$$I_i^{area} = \begin{cases} 1, if\ donor\ catchment\ i\ has\ similar\ area \\ 0, otherwise \end{cases} \quad (1)$$

The original final weight is therefore expressed as Eq. (2).

$$W_i^{final} = W_i^{spatial} \times I_i^{area} \quad (2)$$

Where $W_i^{attribute}$ is derived from either the isotropic or anisotropic displacement formulation. Although this approach accounts for catchment size and displacement location, catchments with similar drainage areas may still respond differently to the same storm because of differences in soil, terrain, land cover, vegetation, and climate. The first incorporates catchment characteristics through unsupervised clustering to identify hydrologically similar donors and adjust their ensemble weights. The second transforms the selected donor hydrographs using a surrogate model to better represent the expected streamflow response of the target catchment. Both approaches are described in detail in the following sections. Beyond the ensemble work, this project developed Replace and Route, a BMI-compliant, model-agnostic component that substitutes a catchment's contribution with discharge from a high fidelity 2D fully distributed model at the hydrofabric nexus, where T-Route carries it downstream. This lets NextGen borrow the skill of models too expensive to run everywhere (See supplementary materials).

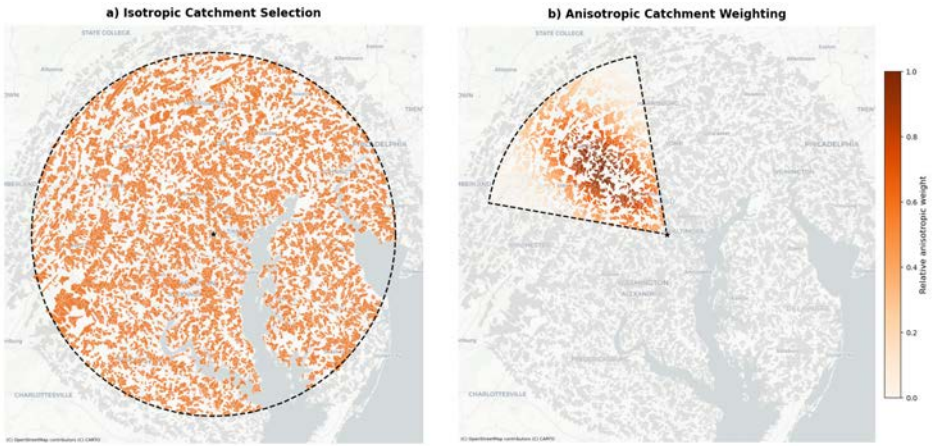


Figure 3. *Isotropic and anisotropic catchment selection around Ellicott City, Maryland.*

4.3.1 Data-Driven Ensemble Refinement

4.3.1.1. Data-driven Attribute-informed donor selection

To improve donor selection, each NextGen hydrofabric catchment is characterized using 49 physiographic and climatic attributes, with detailed descriptions provided in the Supporting Information. Zonal statistics are calculated over each irregular catchment polygon to summarize the underlying gridded datasets at the catchment scale. The attributes represent four broad domains: soil and subsurface properties; land cover and vegetation, including National Land Cover Database (NLCD) [21] including land-cover fractions; terrain and catchment geometry, including drainage area, elevation, and slope; and climate variables such as precipitation, temperature, vapor-pressure deficit,

dew-point temperature, solar radiation, and solar-transmission metrics (Details about data can be found in the SI). After preprocessing and standardization, catchments are grouped using K-means clustering [22] based on their multivariate attribute similarity. The number of clusters is evaluated using the silhouette score [4], elbow method [5], and PCA visualization [6], and four clusters are selected to represent the dominant physiographic catchment groups. An additional attribute-similarity term is then incorporated into both the isotropic and anisotropic ensemble formulations (Eq. (3)).

$$W_i^{final} = W_i^{spatial} \times I_i^{area} \times W_i^{attribute} \quad (3)$$

Where $W_i^{attribute}$ term increases the contribution of donors that are physiographically similar to the target and reduces the contribution of dissimilar donors.

Two approaches are evaluated: Euclidean distance in the normalized, group-weighted 49-dimensional feature space and Euclidean distance in the lower-dimensional t-SNE embedding [7] (SI). In both cases, distance is converted to a similarity weight so that donors closer to the target receive larger weights. A cluster bonus is also applied, donors in the same K-means cluster as the target receive a factor of 1.5, whereas donors in different clusters receive a factor of 1.0. The final weights therefore retain the original spatial logic of the isotropic and anisotropic formulations while incorporating both continuous attribute similarity and discrete cluster membership.

Several weighting configurations are evaluated. C0 (C refers to various NCET configuration) assigns uniform weights, C1 applies isotropic spatial weights, C2 applies anisotropic weights oriented along the estimated displacement vector, and C3 combines anisotropic weighting with the drainage-area filter. C4 and C5 further combine anisotropic weighting with attribute-similarity scores derived from the t-SNE embedding and the standardized 49-dimensional feature space, respectively, together with the cluster bonus. All final weights are normalized to sum to one across the retained donor pool. **Figure 4** presents the resulting donor-weight patterns for each configuration.

4.3.1.2. Surrogate-response ensemble variant

Borrowed hydrographs carry the donor's rainfall-runoff behavior. In this approach, instead of using the direct runoff response, a new rainfall-runoff response is computed using a surrogate model to account for the catchment's physical characteristics and land cover.

For training the model, the NGIAB pipeline was run using a set of 500 synthetic storms generated by Chicago Storm pattern [23] method and CFE model parameters. Each synthetic storm is defined by a set of parameters θ, defined as: T, precipitation length (h); P, total precipitation depth during the event (mm); r, the peak-position ratio (unitless); b and c, shape parameters (unitless); g_w, the initial fractional groundwater storage for the CFE model (unitless); s_m, the initial fractional soil moisture content (unitless); and t_{ini}, initial hour when the storm begins (h). The routed discharge of each catchment [mm h^{-1}] represents the target variable, while the input is the synthetic hyetograph and multiple catchment attributes. The surrogate model is only fed by one dynamic variable (precipitation) and multiple static variables. Due to these characteristics, the selected architecture for the surrogate model is the entity-aware Long Short-Term Memory network (EA-LSTM) [24]. Complementing these inputs, the previous streamflow CFE + Duhamel response is used as the initial cell state value.

The model training split was applied both spatially and by the synthetic storm database. Three zones were delimited to assign the training, validation, and testing, avoiding data leakage. A fourth area was

also identified (around Ellicott City, Maryland) to avoid exposing this data into the machine learning model development and evaluation.

4.3.2 NCET evaluation approach

QPF uncertainty can be expressed in storm-location error, a footprint mismatch, and a different magnitude in the volume and peak intensity of the precipitation. The components of this uncertainty have a major impact in the expected hydrology response. To reduce the uncertainty evaluating NCET, an initial evaluation was done using a fixed footprint of one MRMS event, displacing that footprint to eight cardinal directions with a 100 km displacement distance. The second evaluation was done comparing a real-case of HRRR forecasted precipitation and MRMS estimations for the event of May 27, 2018 for a 18-h forecast issued at 13:00, where a storm-location error was evidenced over Ellicott City headwater catchments.

4.5 Evaluation

The isotropic and anisotropic NCET formulations are evaluated against the deterministic HRRR forecast using the Continuous Ranked Probability Score (CRPS) [25]. CRPS measures how closely the full predictive distribution matches the observed hydrograph while accounting for forecast uncertainty; lower values are better, and zero indicates a perfect forecast. For a deterministic forecast, CRPS reduces to mean absolute error, allowing both forecasts to be compared on the same scale. The relative improvement of each NCET formulation over the deterministic HRRR forecast was quantified using the Continuous Ranked Probability Skill Score (CRPSS), defined as:

$$CRPSS = 1 - \frac{CRPS_{ensemble}}{CRPS_{deterministic}} \quad (4)$$

Positive values indicate improvement over HRRR, zero indicates equal performance, and negative values indicate worse performance.

5. Results

5.1.2 NCET results with MRMS

Four donor-weighting configurations were compared against the MRMS reference: spatial distance only (A), a drainage-area criterion that keeps donors whose area is within a factor of two of the target (B), and two catchment-similarity weights, one from a t-SNE embedding and one from standardized features (C and D). Two cases were tested: the storm on the target catchment, and the storm displaced 100 km in eight directions.

With the storm on the target, the four configurations scored within 2 percent of each other in CRPS (2.98 to 3.03). The t-SNE similarity weight gave the lowest CRPS (2.98), indicating the best performance among the tested configurations. Although the differences were small, t-SNE was the most effective weighting scheme in this scenario. For details about this evaluation, please visit the Hydroshare resource listed in the Supplementary Information section.

Displacing the storm changes this. Every displaced run produced near-zero flow at the target, so the flood relocated rather than disappeared. The isotropic and anisotropic configurations apply the same concentration and differ only in direction: the isotropic weight concentrates on the target and gives the lowest CRPS improvement over an unweighted baseline, while the anisotropic weight concentrates

on the estimated storm location and lowers CRPS substantially across all eight directions. The drainage-area criterion appeared to help but was shown against a random-deletion control to track basin size rather than storm location, and it fails where the storm lands on catchments larger than the target. The similarity weights do not improve on the anisotropic weight once the storm has moved. The anisotropic spatial weight is therefore the best configuration under displacement, and only aiming the weight at the storm's location recovers a displaced flood; drainage-area and attribute similarity do not belong in the spatial weight.

5.1.2 NCET results with HRRR

The two spatial-only NCET configurations, isotropic and anisotropic, were evaluated against the MRMS reference over the 18-hour HRRR forecast window for the three target catchments. Spatial-only means that donor weights are based only on geographic position, without applying drainage-area or catchment-similarity criteria. As shown in **Table 1**, The deterministic HRRR run misses the flood at every catchment, forecasting peaks of 0.3 to 2.1 mm/hr where the MRMS reference peaks range from 20 to 27 mm/hr. Therefore, the point forecast, defined here as the single hydrograph produced by the deterministic HRRR run without an uncertainty range, would have indicated only a minor event.

The isotropic configuration improves on the deterministic forecast, with CRPSS values between +0.29 and +0.30 across the three catchments. However, **Figure 4** shows that the percentile bands remain wide while the median stays near zero. Although the ensemble range contains the MRMS reference, the large spread limits its usefulness. The anisotropic configuration produces substantially greater skill, with CRPSS values of +0.53, +0.57, and +0.60. Its percentile bands are more concentrated around the rising limb, and the median moves closer to the observed peak. The isotropic and anisotropic configurations use the same concentration parameter and differ only in where the weight is directed. This improvement therefore confirms the controlled-experiment finding that aiming the weight toward the estimated storm location is more important than concentration alone.

For cat-114974, the effective ensemble size decreases only modestly under anisotropic aiming, from approximately 172 to 140 donors. Across the three catchments, skill increases slightly with catchment area, from +0.53 at 4.0 km² to +0.60 at 7.0 km². This pattern is consistent with the smoother response expected from larger basins. However, the anisotropic median still underestimates the MRMS peak, showing that spatial reweighting alone does not fully reproduce the target flood response. This remaining difference motivates the surrogate-response approach presented in the following section, in which the donor hydrographs are corrected rather than only reweighted.

Table 1. *NCET performance on the 18-hour HRRR forecast against the MRMS reference.*

Catchment	Area (km²)	Reference peak (mm/hr)	HRRR det. peak (mm/hr)	CRPS det.	CRPS iso	CRPS aniso	$CRPS_{ss}$ iso	$CRPS_{ss}$ aniso
cat-114975	4.0	27.10	1.33	5.171	3.620	2.454	+0.300	+0.525
cat-114974	5.6	26.02	0.31	5.036	3.534	2.147	+0.298	+0.574
cat-114976	7.0	20.58	2.10	3.908	2.760	1.556	+0.294	+0.602

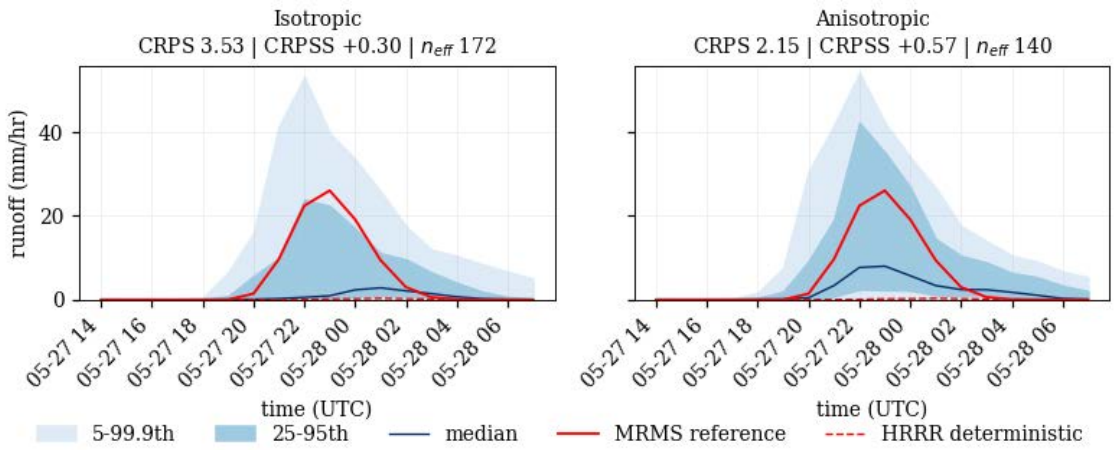


Figure 4. *Isotropic and anisotropic NCET forecasts for cat-114974 (5.6 km²) over the 18-hour HRRR forecast period. Shaded areas show the ensemble percentile ranges, the blue line is the ensemble median, and the red lines represent the MRMS reference and deterministic HRRR forecast. CRPS measures forecast error, CRPSS shows improvement relative to HRRR, and neff is the effective ensemble size.*

5.2.2 Surrogate-response ensemble variant

The model was trained for 18 epochs, where the lowest validation loss (Mean Square Error) was achieved at epoch 8. The Nash-Sutcliffe Efficiency (NSE) metric for the testing dataset was 0.83. This metric penalizes high flow value mismatches that, for a flash flood surrogate model, suits with the surrogate model evaluation. The performance obtained is widely accepted as a good performance [26].

The surrogate model was run for the donor catchments and weights evaluated in Section 5.2.1. A set of predefined precipitation scale levels was also applied (0.6× to 2.8×) to evaluate the QPF-QPE magnitude sensitivity and mismatch. The results for two different approaches are presented in **Figure 5**: (1) using only the results of the surrogate model; and (2) mixing the results of the surrogate model with NCET outputs.

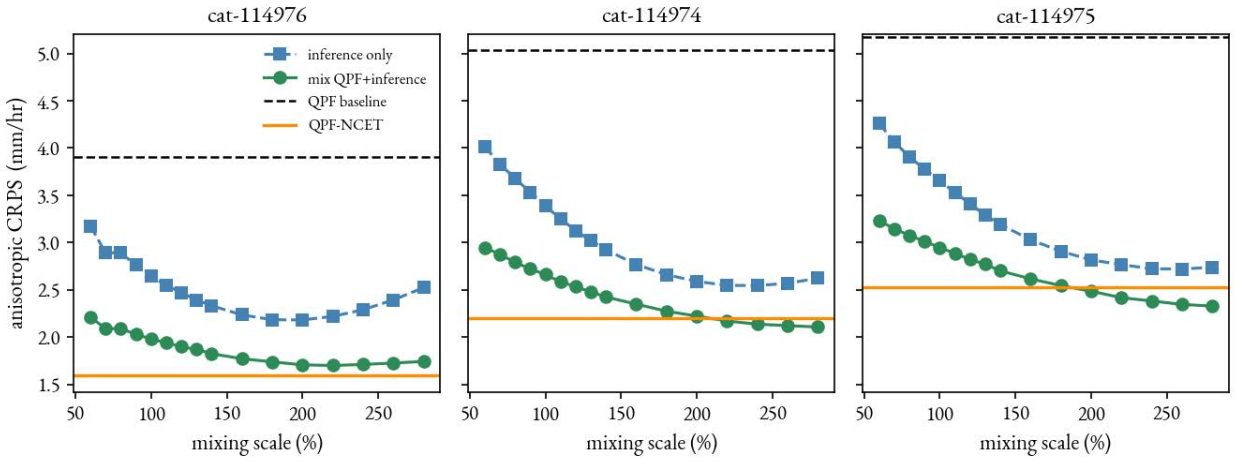


Figure 5. *CRPS surrogate-response for the target catchments. The x-axis represents the scaling factor of the surrogate model's input precipitation, and the y-axis represents the metric CRPS. The surrogate-only results (blue markers) and a mixed response with the QPE results (green markers) are presented. The QPF baseline is represented by a black dotted line, while the anisotropic NCET, by an orange line.*

The behavior of the surrogate model's responses indicates that there is a mismatch between the QPF and the QPE products. Using the same NCET approach applied to precipitation to create a weighted mean hyetograph for two scale factors (1.0× and 2.2×), for the isotropic and anisotropic approaches (**Figure 6**), the results indicate that QPE and QPF storms have different intensities, shapes, and extensions.

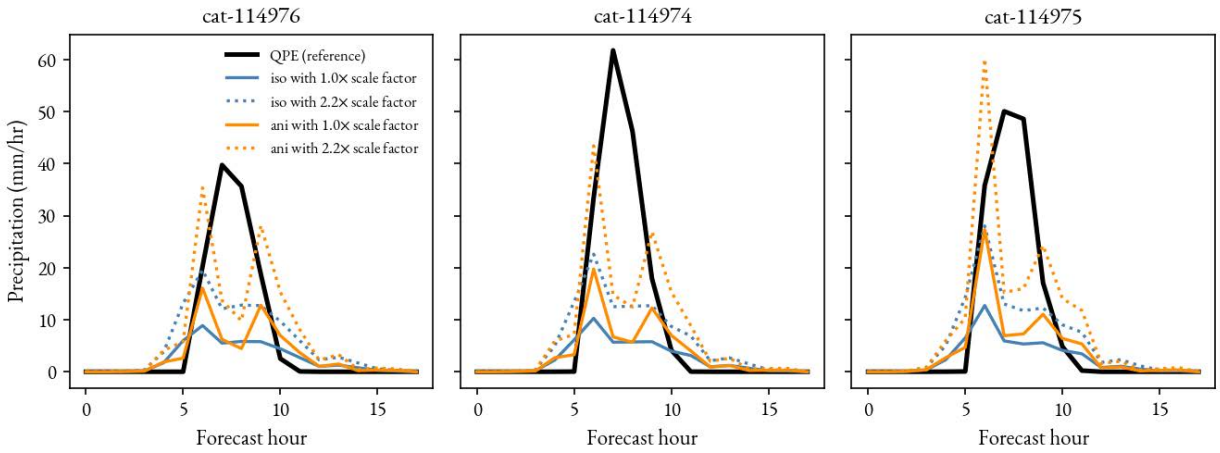


Figure 6. *Comparison between the QPE hyetograph of the target catchment and the weighted mean hyetographs for 1.0× and 2.2× precipitation scales generated from NCET weights for the isotropic (iso) and anisotropic (ani) approaches.*

The scaling factor corrects the mismatch between the two products partially, although there are two main points to consider: (1) The surrogate model was trained with a Chicago storm generator following a type-III-like storm with randomized parameters. This storm type matches the QPE hyetograph (**Figure 5**), but the QPF hyetograph has a bimodal shape which translates into two runoff pulses in the target catchments. (2) Based on the difference in both products' hyetographs, the scaling factor helps to adjust the peak precipitation and accumulated depth during the event.

6. Conclusion

This project adapted NPET to the catchment-based NextGen hydrofabric and evaluated the framework using both a controlled storm-displacement experiment and a real HRRR forecast for the May 27, 2018 Ellicott City, Maryland, flash-flood event. Catchment similarity improved the ensemble when the storm remained centered over the target catchment. Once the storm was displaced, however, neither physiographic similarity nor the drainage area criterion consistently improved the ensemble because both could downweight donor catchments carrying the displaced flood signal. Under these conditions, the key factor was not how similar a donor catchment was to the target, but whether the spatial weighting was aligned with the displaced storm location. The largest improvement was obtained using an anisotropic weighting function directed toward the estimated storm-displacement vector, whereas an isotropic weighting function with comparable concentration but centered on the target produced only limited benefit. These results indicate that storm-location information should play the primary role in guiding spatial weights when forecast precipitation is spatially displaced. Although the surrogate model exhibited a better CRPS, the current limitation is the training database, which includes only one storm type. Adding more storm types or shapes will improve the model's performance. In parallel, remapping the precipitation to match the QPE product could improve the performance, but adds an extra complexity level towards NCET operational usage.

Acknowledgments: The authors gratefully acknowledge Dr. Fred Ogden (Chief Scientist, NOAA-NWS Office of Water Prediction) for the insightful guidance about the NextGen framework and the National Water Model, his talks were key for comprehending NGIAB modules and their interaction. We also acknowledge Quinn Russell, Dongha Kim, and the team of the Alabama Water Institute for their technical support, troubleshooting and explanation about T-Route implementation in the NextGen framework. We would like to extend our deepest gratitude to our theme leads, Dr. Humberto Vergara and Dr. Mohamed Abdelkader, for their unwavering support. Likewise, we would like to express our gratitude to our academic advisors, Dr. Andrew Bennett (The University of Arizona), Professor Reza Khanbilvardi (The City College of New York), Professor Hamid Moradkhani (The University of Alabama), Dr. Mojtaba Sadegh (Boise State University), whose broader support and encouragement have contributed to our academic development. We sincerely thank Dr. Megan Vardaman and Nana Oye Djan for their insightful comments, suggestions, and above all, for their tireless support and organization of the Summer Institute.

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA.

The authors used a large language model (Anthropic's Claude) to assist with grammatical editing and coding. All factual content, results, and interpretations were reviewed and verified by the authors.

Supplementary Materials:

1. https://github.com/NWC-CUAHSI-Summer-Institute/NCET-data-preprocess
2. https://www.hydroshare.org/resource/96f00a9eaef44b3ba4c00ddb24f70087/

References

[1] G. Terti, I. Ruin, S. Anquetin, and J. J. Gourley, "A situation-based analysis of flash flood fatalities in the United States," *Bull. Am. Meteorol. Soc.*, vol. 98, no. 2, pp. 333–345, 2017.

[2] M. Špitalar, J. J. Gourley, C. Lutoff, P.-E. Kirstetter, M. Brilly, and N. Carr, "Analysis of flash flood parameters and human impacts in the US from 2006 to 2012," *J. Hydrol.*, vol. 519, pp. 863–870, 2014.

[3] A. Ahmadalipour and H. Moradkhani, "A data-driven analysis of flash flood hazard, fatalities, and damages over the CONUS during 1996–2017," *J. Hydrol.*, vol. 578, p. 124106, 2019.

[4] L. Cuo, T. C. Pagano, and Q. J. Wang, "A review of quantitative precipitation forecasts and their use in short-to medium-range streamflow forecasting," *J. Hydrometeorol.*, vol. 12, no. 5, pp. 713–728, 2011.

[5] H. Yan and W. A. Gallus Jr, "An evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain," *Weather Forecast.*, vol. 31, no. 4, pp. 1363–1379, 2016.

[6] H. Wernli, M. Paulat, M. Hagen, and C. Frei, "SAL—A novel quality measure for the verification of quantitative precipitation forecasts," *Mon. Weather Rev.*, vol. 136, no. 11, pp. 4470–4487, 2008.

[7] W. F. Krajewski, R. Goska, R. Post, F. Quintero, and N. Velasquez, "Is This Rainfall Forecast Good or Bad? For Flood Forecasting, the Answer Is Scale Dependent," *Bull. Am. Meteorol. Soc.*, vol. 106, no. 9, pp. E1772–E1793, 2025.
[8] H. Vergara, J. J. Gourley, and M. Erickson, "An efficient ensemble technique for hydrologic forecasting driven by quantitative precipitation forecasts," *J. Hydrometeorol.*, vol. 24, no. 3, pp. 479–495, 2023.
[9] Z. P. Brodeur, C. Delaney, B. Whitin, and S. Steinschneider, "Synthetic forecast ensembles for evaluating forecast informed reservoir operations," *Water Resour. Res.*, vol. 60, no. 2, p. e2023WR034898, 2024.
[10] R. A. Clark, J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, "CONUS-wide evaluation of National Weather Service flash flood guidance products," *Weather Forecast.*, vol. 29, no. 2, pp. 377–392, 2014.
[11] T. L. Sweeney, "Modernized areal flash flood guidance," 1992.
[12] J. J. Gourley *et al.*, "The FLASH project: Improving the tools for flash flood monitoring and prediction across the United States," *Bull. Am. Meteorol. Soc.*, vol. 98, no. 2, pp. 361–372, 2017.
[13] J. M. English *et al.*, "Evaluating operational and experimental HRRR model forecasts of atmospheric river events in California," *Weather Forecast.*, vol. 36, no. 6, pp. 1925–1944, 2021.
[14] B. Roberts *et al.*, "What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms?," *Weather Forecast.*, vol. 35, no. 6, pp. 2293–2316, 2020.
[15] D. C. Dowell *et al.*, "The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description," *Weather Forecast.*, vol. 37, no. 8, pp. 1371–1395, 2022.
[16] E. P. James *et al.*, "The High-Resolution Rapid Refresh (HRRR): an hourly updating convection-allowing forecast model. Part II: Forecast performance," *Weather Forecast.*, vol. 37, no. 8, pp. 1397–1417, 2022.
[17] W. A. Gallus Jr, A. Duhachek, K. J. Franz, and T. Frazier, "A Climatology of Errors in HREF MCS Precipitation Objects," *Water*, vol. 17, no. 15, p. 2168, 2025.
[18] F. L. Ogden *et al.*, "The NextGen water resources modeling Framework: Community innovation at the intersection of hydrologic, data and computer sciences," *JAWRA J. Am. Water Resour. Assoc.*, vol. 62, no. 1, p. e70089, 2026.
[19] A. Patel *et al.*, "NextGen In A Box (NGIAB): Open-Source containerization of the NextGen framework to enable community-driven hydrology modeling," *Environ. Model. Softw.*, p. 106666, 2025.
[20] S. A. Koriche *et al.*, "Enhancing Hydrologic Process Representation in the CFE Model through Nash-Cascade Reservoirs: A Comparative Analysis of CFE 1.0 and CFE 2.0," in *AGU Fall Meeting Abstracts*, 2024, vol. 2024, no. 769, pp. H31L-0769.
[21] J. Dewitz, "National land cover database (NLCD) 2019 products (ver. 3.0, february 2024)," *US Geol. Surv. Data Release*, p. 624, 2021.
[22] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, 2020, pp. 747–748.
[23] C. J. Keifer and H. H. Chu, "Synthetic storm pattern for drainage design," *J. Hydraul. Div.*, vol. 83, no. 4, pp. 1331–1332, 1957.
[24] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing, "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets," *Hydrol. Earth Syst. Sci.*, vol. 23, no. 12, pp. 5089–5110, 2019.

[25] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.

[26] D. N. Moriasi, M. W. Gitau, N. Pai, and P. Daggupati, "Hydrologic and water quality models: Performance measures and evaluation criteria," *Trans. ASABE*, vol. 58, no. 6, pp. 1763–1785, 2015.

Chapter 3: Sensitivity of Flash Flood Predictions to Precipitation Forcing Uncertainty for Different Spatiotemporal Resolutions

Bikas Chandra Gupta[1], Alexander Joseph Lastner[2], Danna Villarreal[3], Mohamed Abdelkader[4*], and Humberto Vergara[5*]

[1]University of Texas, Arlington; *bikaschandra.gupta@uta.edu*
[2]University of Maryland, College Park; *alastner@umd.edu*
[3]University of Arkansas; *dannav@uark.edu*
[4]University of Iowa; *mohamed-abdelkader@uiowa.edu*
[5]University of Iowa; *humberto-vergaraarrieta@uiowa.edu*
[*]Theme Leader

Abstract: Accurate flash flood prediction is dependent on both the quality of the hydrological model and the precipitation forcing inputs that drive it. The Next Generation Water Resources Modeling Framework (NextGen) is model-agnostic, potentially enabling model formulations fitted to flash flood forecasting. Flash flooding, however, requires sub-hourly updates of both the water balance and the routing, and the ability to test model sensitivity at those scales remains limited. This research established a pipeline for quantifying the sensitivity of actionable flash flood predictions to precipitation forcing uncertainty and spatiotemporal resolution. Three historical case study flash floods, one in North Carolina, Pennsylvania, and Texas, were simulated with Conceptual Functional Equivalent (CFE) and T-Route modules in NextGen In A Box (NGIAB). NGIAB was modified to run on a 15-min timestep and forced with 15-minutes Multi-Radar Multi-Sensor (MRMS) precipitation. By iteratively modifying precipitation magnitude scaling, spatial resolution, and temporal resolution, 3-dimensional error matrices were populated for peak flow magnitude bias and timing lag. These matrices were then converted to 2-dimensional matrices that quantified the minimum uncertainty in precipitation input necessary to produce predictions with ≥5% probability of having a peak flow magnitude bias $\geq|50\%|$ and a lag time $\geq|0.5\ \text{hr}|$. Low sensitivity of critical uncertainty to increases in spatiotemporal resolution were observed beyond ~7 km and ~1 hr. This was especially pronounced for lag time matrices, with critical uncertainty being insensitive to precipitation forcing uncertainty at resolutions beyond the temporal threshold and critical uncertainties near zero (i.e., simulations did not produce acceptably low lag time predictions at resolutions without uncertainty) at resolutions below the threshold. The general insensitivity to scale compared to previous studies may be due to the aggregation approach smearing rather than 'missing' zones of highly localized intense precipitation due to spare observations, resulting in total storm precipitation volume being conserved and in a reduced impact of basin rainfall volumes. Future development of this pipeline should focus on investigating additional basins, controlling for the effects of storm idiosyncrasies in evaluating sensitivity, and exploring additional uncertainty definitions and unactionable prediction error thresholds.

1. Motivation

Floods are the second most deadly weather hazard across the United States [1]. Flash floods, or floods that occur within six hours of a triggering event, such as intense rains [2], are especially dangerous. Between 1996 and 2017, the National Weather Service (NWS) recorded 74,814 flash floods, resulting in 1,399 fatalities and $75.6 billion in damages [3]. Over 99% of these events were driven by intense rainfalls [3], with patterns in flash flood timing and intensity reflecting the spatial variability of hydroclimatology across CONUS [2]-[4]. Although exceptionally severe flash floods, such as those caused by dam failures, have killed hundreds in a single event, most deaths from flash floods are caused by events that individually result in fewer than five deaths [5]. The ability to accurately predict even minor pluvial flash floods and communicate their risks, then, may greatly reduce the number of total flash flood fatalities.

Accurate and timely flash flood prediction is dependent on both the accuracy of the models used to predict streamflow and the accuracy and timeliness of the precipitation forcing inputs ingested into those models. There is no universal hydrological model that is adequate for global flood prediction [6], and model selection is often driven by legacy rather than adequacy for the problem and location of interest [7]. Rather than advancing any one model, the Next Generation Water Resources Modeling Framework (NextGen) provides a framework that increases the feasibility of running many models and promotes model experimentation [8]. NextGen In A Box (NGIAB) is a containerized distribution that simplifies the implementation of the framework for hydrological prediction and model experimentation [9]. Forcing for NextGen, however, is currently limited to NOAA Analysis of Record for Calibration (AORC) [10] data and hardcoded to operate only at hourly resolutions, narrowing the scope of potential NextGen-enabled model experiments for model sensitivity to precipitation forcing uncertainty or spatiotemporal resolution of meteorological forcing. Effectively leveraging NGIAB to conduct sensitivity analyses at sub-hourly temporal resolutions requires the creation of a modified NGIAB instance.

2. Objectives and Scope

This research aimed to establish a pipeline for investigating the sensitivity of actionable (i.e., consistently useful) flash flood predictions to precipitation forcing uncertainty and spatiotemporal resolution in NGIAB through a case study of three historical flash flood events. This required the integration of non-AORC precipitation data at sub-hourly resolutions into NGIAB. This study also developed a strategy to express whether a prediction is actionable, meaning its error remains within defined tolerances, given the precipitation forcing uncertainty and spatiotemporal resolution.

3. Previous Studies

Concerns about the impact of precipitation forcing uncertainty and spatiotemporal resolution are frequently addressed in hydrological modeling research; however, these studies have not considered the effects of forcing uncertainty, temporal resolution, and spatial resolution in tandem. An investigation of runoff model sensitivity to radar rainfall resolution for two semi-arid watersheds found that the convergence of runoff and discharge volume predictions with increasing rainfall resolution only occurred due to the convergence of the resolution of precipitation forcing with the basin simulation resolution [11]. Other studies have found that, although higher spatial resolution precipitation observations may improve rainfall volume estimates and improve predictions, finer spatial resolution of heterogeneity in precipitation occurrence without improved volumetric estimates have a negligible effect on estimations of primary peak flows [12]-[13]. In contrast, higher temporal resolution precipitation forcings are likely to improve model predictions if precipitation volume for

an event is accurately resolved [13]. Flash flood-specific research corroborates these inferences, finding diminishing returns with increasing spatial resolution but improved predictive power with increasing temporal resolution until forcing resolution exceeds that of streamflow observation [14]. Another study investigating the sensitivity in flash flood predictions to uncertainty in rainfall estimates, watershed model parameters, and initial soil moisture conditions for semiarid basins found the spatial and volumetric precipitation uncertainty to be the largest single parameter contributor to total source uncertainty, although potentially rivaled by certain parameters representing hillslope characteristics [15].

4. Methodology

The pipeline comprised the following steps: the identification of flash flood scenarios, including both basin and storm events, that are compatible with available forcing data; the running of baseline and modified flash flood scenarios using NGIAB modified to run on 15-min timesteps; and the computation of forcing uncertainty thresholds on the actionability of predictions using conditional probabilities derived from the simulations. This pipeline is summarized in **Figure 1**.

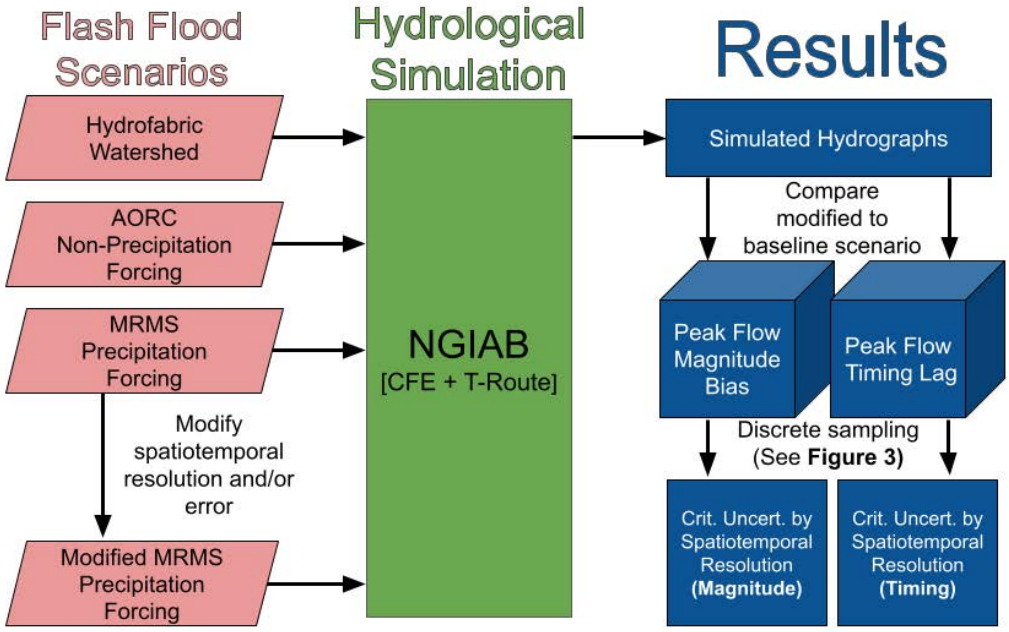


Figure 1. *Methodological pipeline for evaluating flash flood prediction sensitivity to precipitation forcing error or spatiotemporal resolution.*

4.1. Event and Site Selection

Three flash flood-generating pluvial events from the contiguous United States were investigated: greater Asheville, North Carolina (NC); greater Pittsburgh, Pennsylvania (PA); and greater Kerrville; Texas (TX; see **Figure 2**).

Figure 2. *Geolocation of the three case study sites along with Hydrofabric catchments used for NextGen flash flood simulations.*

Flood modelling was performed for select catchments that satisfy the following criteria: (1) experienced a flash flood resulting from the storm event identified for the scenario, (2) classified as a GAGES-II reference basin [16]-[17], and (3) possessed a drainage area <1000 km^2. The following USGS gauges and simulation periods corresponding to the selected catchments are summarized in **Table 1**.

Table 1. *Summary of historical flash flood events and NextGen simulation windows for the three case study catchments.*

State (USGS Gage)	Hydrofabric ID	Drainage Area (km^2)	Observed Peak Flow (m^3/s)	Peak Date & Time (UTC)	Simulation Window
NC (03450000)	1016883	14	226	Sept 27, 2024 (14:00)	Sept 24-30, 2024
PA (03049800)	807634	15	51	Apr 12, 2024 (01:30)	Apr 09-15, 2024
TX (08165500)	2414792	746	8920	July 4, 2025 (10:00)	July 01-07, 2025

4.2. Data Sources

Meteorological and spatial data utilized in this study are summarized in **Table 2**. The Hydrofabric served as the geographic structural basis of the model. To drive the simulations, AORC was used to establish baseline non-precipitation forcing conditions, and MRMS quantitative precipitation estimates (QPE) provided the high-resolution (15-min) event-based rainfall forcing required to capture flash flood dynamics.

Table 2. *Datasets and metadata utilized for model configuration and meteorological forcing.*

Dataset	Resolution	Counts/Frequency	Availability	Provider
Hydrofabric	Vector	~0.8 million elements	2022	NOAA [18]
AORC	1-km	1 hr	1979-present	NOAA [10]
MRMS QPE	1-km	2-min	2012-present	NOAA [19]-[21]

4.3. Simulation Scenario Design

A comprehensive sensitivity analysis was performed by iteratively simulating flood events across a full-factorial grid (2,280 individual simulation scenarios per case study basin) of precipitation forcing parameters as summarized in **Table 3**.

Table 3. *Experimental design parameters for the 2,280-scenario sensitivity analysis per case study basin.*

Parameter	Baseline Value	Range	Step Size
Precipitation Magnitude Scaling	1.0x	0.1x – 1.9x	0.1x
Spatial Resolution	1 km	1 – 10 km	1 km
Temporal Resolution	15 min	15 – 180 min	15 min

4.4. Hydrological Simulation within Containerized NextGen Framework

Within NGIAB, the modules used in this analysis were the Conceptual Functional Equivalent (CFE) model and T-Route, which represent a rainfall–runoff hydrologic model and a dynamic channel routing model for streamflow routing, respectively [8]-[9]. A new model realization was first generated using the "NGIAB_data_preprocess" python package (https://github.com/CIROH-UA/NGIAB_data_preprocess [9]), which automatically created the configuration files required to execute the model simulations. These files include the "*realization.json*", which defines the overall simulation settings and model components, "*troute.yaml*", which configured the T-Route routing model, and the NOAH-OWP-M catchment configuration files, which specify the numerical settings for the water balance model. These configuration files were modified to use a 15-min time step or its equivalent in the required units, to ensure that the hydrologic and routing components operated correctly for the desired simulations. Additionally, a new Docker image was created to execute NGIAB with the modified components as part of the sub-hourly simulation workflow.

4.5. Prediction Sensitivity to Precipitation Forcing Uncertainty and Spatiotemporal Resolution

The simulated peak flow magnitude and timing of these modified forcing scenarios were compared to the unmodified baseline simulation to compute magnitude bias and temporal lag, respectively. By performing this comparison iteratively for all combinations of considered magnitude scaling, temporal resolution, and spatial resolution, a 3-dimensional matrix of output departure from baseline (considered error in this study) was populated for both peak flow magnitude bias and timing lag.

These 3-dimensional error matrices were then converted to 2-dimensional uncertainty response matrices through iterative discrete sampling for each spatiotemporal resolution (i.e., specific combination of spatial and temporal resolutions). Forcing uncertainty was represented as a scaling of precipitation magnitude, expressed as the absolute percent error that contains 95% of actual forcing error (e.g., for an uncertainty of 25%, the probability that magnitude scaling falls between 0.75 and 1.25 is 95%). Scalings were assumed to be normally distributed around 1. Critical uncertainty is the level of uncertainty on precipitation forcing scaling magnitude that results in ≥5% of predictions becoming unactionable, providing a measure of tolerance for precipitation forcing uncertainty. Acceptable prediction error thresholds are infrequently discussed in hydrological modeling literature, but thresholds should be fit to purpose [22]. Whether a warning is necessary depends on peak magnitude, and the timing of a warning should depend on the peak timing. As a result, the two do not share tolerance thresholds. Tolerances are also rarely symmetric (e.g., a flash flood occurring an hour earlier than predicted will likely be more dangerous than one occurring an hour later). Dedicated

research on effective thresholds is warranted but is beyond the scope of this work; therefore, symmetric bias thresholds of $\geq|50\%|$ for peak flow and $\geq|0.5\ \text{hr}|$ for lag on peak timing were adopted. Uncertainty was then incrementally increased from 0.1%-90% in steps of 0.1% to generate Gaussian distributions. Each distribution was discretized to estimate the sampling weight of flood errors by the forcing scalings that produced them. Iteration continued until critical uncertainty was achieved. The procedure was repeated for every spatiotemporal resolution defined to generate the 2-dimensional uncertainty response matrices. A schematic representation of this process appears in **Figure 3**.

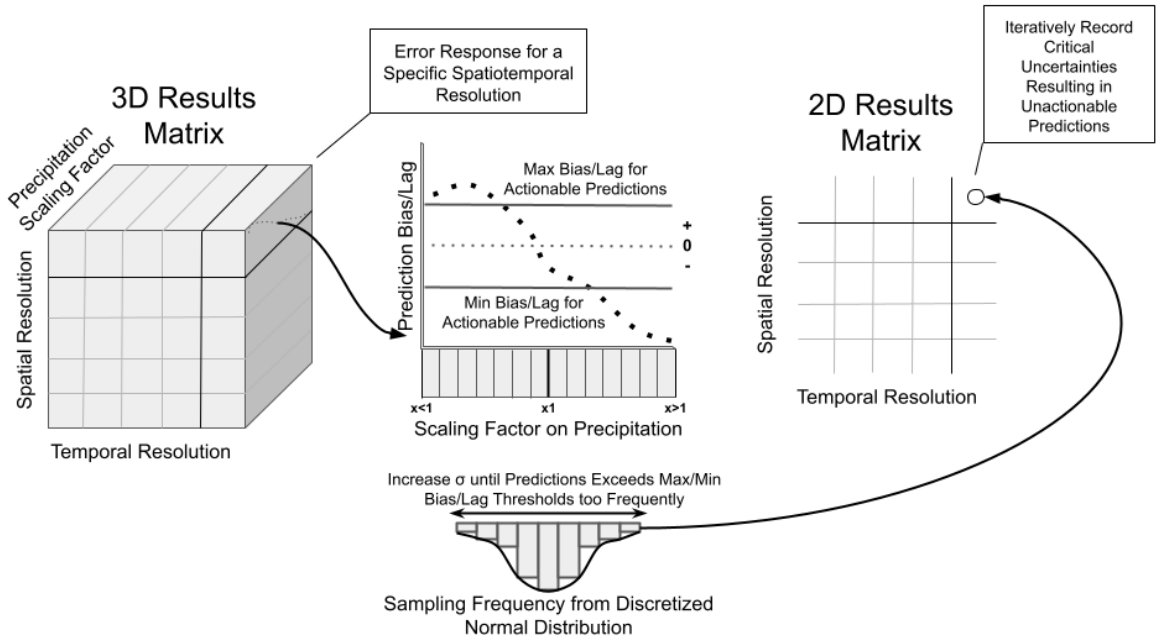


Figure 3. *2-dimensional Critical uncertainty estimation from 3-dimensional results matrix procedure. σ (variance) = precipitation forcing uncertainty / 1.96.*

The reduction to two dimensions integrates over the magnitude scaling axis, measuring how much forcing error a prediction can tolerate; however, it cannot attribute that variability to individual factors. The same 3-dimensional matrices were therefore also analyzed with an analysis of variance (ANOVA) decomposition of the total sum of squares (SST) [23] to quantify the relative contributions of precipitation magnitude scaling, spatial resolution, temporal resolution, and their interactions to the variability in simulated peak flow magnitude and timing. Because the decomposition does not depend on the actionability thresholds, it also serves as a threshold-free check that the patterns in the critical uncertainty matrices are not an artifact of the chosen limits. Results are reported in the **Supplementary Material**.

5. Results

Critical uncertainty responds to resolution in discrete steps (**Figure 4**). Increasing spatial resolution to finer than ~7 km and temporal resolution than ~1 hr did not substantially impact critical uncertainty. Critical uncertainties for NC, PA, and TX, respectively, for resolutions finer than 7 km and 1 hr were 40%, 25%, and 40% for peak magnitude and 90%, 30%, 85% for temporal lag. At coarser resolutions, values were closer to 30%, 20%, 30% for peak magnitude. For NC and TX, critical uncertainty for lag sits near the top of the resolvable range when the timestep is shorter than about 2.5 hr, and falls to

near zero when it is longer. At those longer timesteps the model did not produce acceptable lag predictions even with no forcing error. Narrow bands of near-zero critical uncertainty for lag also appear at timesteps up to about 1 hr, against that trend. They do not follow the resolution pattern and are most likely the result of aliasing or a similar artifact of the sampling and aggregation procedure. Forcing at a finer timestep, for example 2-min MRMS, would test this.

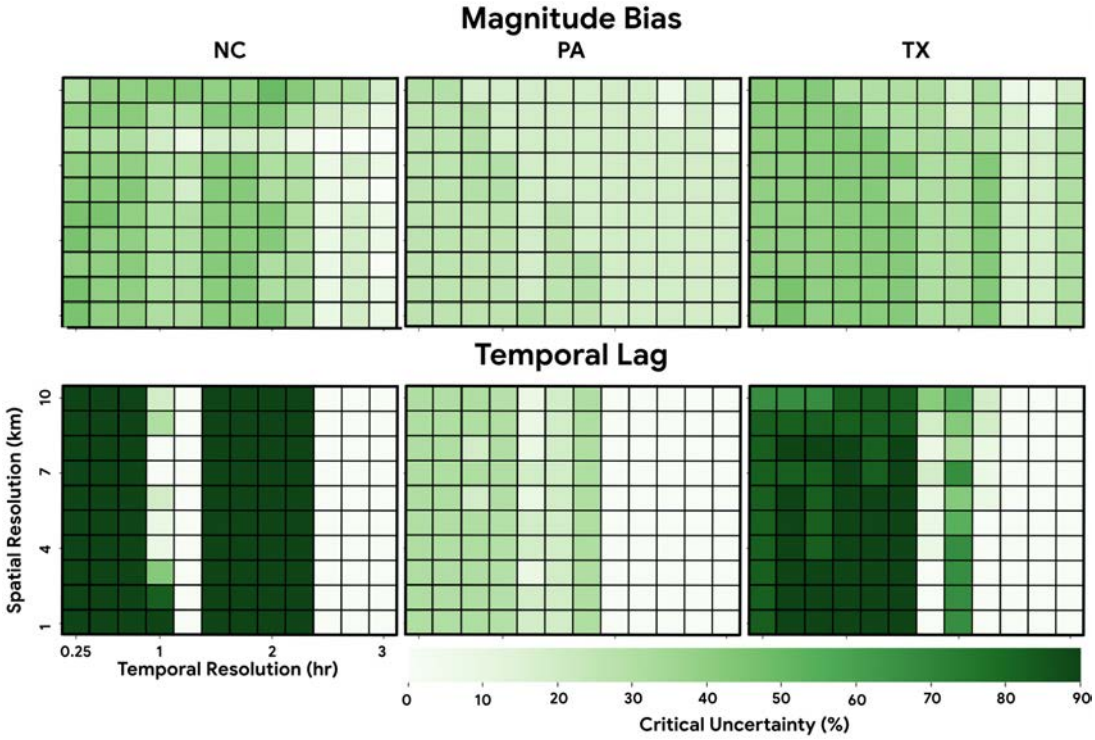


Figure 5. *Critical uncertainty matrices for NC, PA, and TX flash floods. Columns correspond to the same location, and rows to the same flood variable (i.e., magnitude and temporal lag). Critical uncertainty values of values below 0.1% and above 90% were rounded to the limits of the colorbar and correspond to unresolvably low and high critical uncertainties for the simulation procedure, respectively.*

Spatial aggregation had little effect because it smeared intense localized precipitation rather than missing it. Total storm volume was conserved, resulting in the rainfall delivered to each basin changing only at the coarsest grid spacings. Temporal aggregation lowers the peak intensity of the forcing, which damps the flashiness of the simulated response (**Figure 5**). That damping also offers an explanation for the anomalous bands. When forcing intensity is reduced, a secondary peak can exceed the primary one, and the reported time to peak jumps to a different part of the hydrograph. This would produce isolated bands of near-zero critical uncertainty at otherwise fine timesteps. At timesteps longer than about 2.5 hr, peak timing cannot be resolved within a one-hour window without additional information, which explains the collapse of critical uncertainty there. A variance decomposition of the same 3-dimensional matrices supports this reading, and does so without reference to the actionability thresholds. It is reported in the **Supplementary Material**.

6. Conclusion

This research suggests that the sensitivity of flash flood magnitude and timing predictions to precipitation forcing uncertainty responds only modestly to improvements in spatiotemporal resolution beyond some critical threshold, assuming the volume of precipitation is unaffected by

sampling resolution. For the case study flash flood scenarios investigated, spatial and temporal resolutions exceeding 7 km and 1 hr, respectively, are likely to provide diminishing returns when predicting peak flash flood magnitudes with biases < |50%| and peak flood timing within <|0.5 hr|.

Although this research successfully developed and tested a pipeline for evaluating flash flood prediction sensitivity to precipitation forcing uncertainty and spatiotemporal resolution, the scope and pipeline methodology can be further improved. In addition to evaluating additional observed flash flood scenarios, precipitation data for observed flash floods could be compiled into a catalog and then applied iteratively as precipitation forcing to new basins, potentially controlling for storm idiosyncrasies during sensitivity analysis. The generation of synthetic storms using a weather generator (e.g., RainyDay [24]) may also suffice. The relationship between patterns in basin critical uncertainty matrices and GAGES-II basin characteristics could then be evaluated. Additionally, alternative resolutions, models, uncertainty definitions and distributions, and unactionable prediction threshold definitions could be investigated in future iterations of this research to better align with actual forcing uncertainties and the objectives of operational flash flood prediction.

Acknowledgements: The authors wish to especially acknowledge support and assistance from Aldo Tapia Araya (University of Arizona, Summer Institute Fellow), Josh Cunningham (Alabama Water Institute), Nana Oye Djan (Carnegie Mellon University, Summer Institute Coordinator), Quinn Lee (Alabama Water Institute), Nia Minor (Alabama Water Institute), Arpita Patel (Alabama Water Institute), and Dr. Megan Vardaman (Summer Institute Coordinator). The authors also wish to thank their academic advisors: Dr. Rebecca L. Muenich (University of Arkansas), Dr. Karen L. Prestegaard (University of Maryland), and Dr. Adnan Rajib (University of Texas Arlington).

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA.

This research utilized CIROH Cyberinfrastructure external resources - NSF ACCESS allocation number EES240087 on Indiana University's Jetstream2 managed by CIROH Cyberinfrastructure. ACCESS is supported by NSF awards #2138259, #2138286, #2138307, #2137603, and #2138296. The authors appreciate support from the CIROH Cyberinfrastructure team. Learn more: https://hub.ciroh.org/docs/services/intro. See references [25]-[26] for more details.

Gemini 2.5, Gemini 3.5, and GPT-5.5 (June-July, 2026) aided in coding and debugging. All factual content and interpretations were verified by the authors.

Supplementary Materials: The computational framework, analysis scripts, and model configuration files used in this study are archived and publicly available in the following GitHub repository: https://github.com/NWC-CUAHSI-Summer-Institute/nextgen_spatiotemporal_sensitivity.

Additional details on the Total Sum of Squares (SST) methodology, results, and illustrative figures are available in the HydroShare repository: http://www.hydroshare.org/resource/bab353ccc6f74c6d8a355c4b096f7371.

References

[1] National Weather Service, Summary of Natural Hazard Statistics for 2024 in the United States. Silver Spring, MD, USA: National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 2025. [Online]. Available: https://www.weather.gov/media/hazstat/sum24.pdf

[2] M. Saharia, P.-E. Kirstetter, H. Vergara, J. J. Gourley, and Y. Hong, "Characterization of floods in the United States," Journal of Hydrology, vol. 548, pp. 524–535, 2017. doi: 10.1016/j.jhydrol.2017.03.010.

[3] A. Ahmadalipour and H. Moradkhani, "A data-driven analysis of flash flood hazard, fatalities, and damages over the CONUS during 1996–2017," Journal of Hydrology, vol. 578, Art. no. 124106, 2019. doi: 10.1016/j.jhydrol.2019.124106.

[4] M. Saharia, P. Kirstetter, H. Vergara, J. J. Gourley, Y. Hong, and M. Giroud, "Mapping flash flood severity in the United States," Journal of Hydrometeorology, vol. 18, pp. 397–411, 2017. doi: 10.1175/JHM-D-16-0082.1.

[5] S. T. Ashley and W. S. Ashley, "Flood fatalities in the United States," Journal of Applied Meteorology and Climatology, vol. 47, pp. 805–818, 2008. doi: 10.1175/2007JAMC1611.1.

[6] G. Blöschl et al., "Twenty-three unsolved problems in hydrology (UPH)—A community perspective," Hydrological Sciences Journal, vol. 64, no. 10, pp. 1141–1158, 2019. doi: 10.1080/02626667.2019.1620507.

[7] N. Addor and L. A. Melsen, "Legacy, rather than adequacy, drives the selection of hydrological models," Water Resources Research, vol. 55, no. 1, pp. 378–390, 2019. doi: 10.1029/2018WR022958.

[8] F. L. Ogden et al., "The NextGen Water Resources Modeling Framework: Community innovation at the intersection of hydrologic, data and computer sciences," Journal of the American Water Resources Association, vol. 62, no. 1, Art. no. e70089, 2026. doi: 10.1111/1752-1688.70089.

[9] A. Patel et al., "NextGen In A Box (NGIAB): Open-source containerization of the NextGen framework to enable community-driven hydrology modeling," Environmental Modelling & Software, vol. 193, Art. no. 106666, 2025. doi: 10.1016/j.envsoft.2025.106666.

[10] G. Fall et al., "The Office of Water Prediction's Analysis of Record for Calibration, Version 1.1: Dataset description and precipitation evaluation," Journal of the American Water Resources Association, vol. 59, no. 6, pp. 1246–1272, 2023. doi: 10.1111/1752-1688.13143.

[11] F. L. Ogden and P. Y. Julien, "Runoff model sensitivity to radar rainfall resolution," Journal of Hydrology, vol. 158, nos. 1–2, pp. 1–18, 1994. doi: 10.1016/0022-1694(94)90043-4.

[12] C. Obled, J. Wendling, and K. Beven, "The sensitivity of hydrological models to spatial rainfall patterns: An evaluation using observed data," Journal of Hydrology, vol. 159, nos. 1–4, pp. 305–333, 1994. doi: 10.1016/0022-1694(94)90263-1.

[13] Y. Huang, A. Bárdossy, and K. Zhang, "Sensitivity of hydrological models to temporal and spatial resolutions of rainfall data," Hydrology and Earth System Sciences, vol. 23, pp. 2647–2663, 2019. doi: 10.5194/hess-23-2647-2019.

[14] A. Rafieeinasab et al., "Toward high-resolution flash flood prediction in large urban areas—Analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic modeling," Journal of Hydrology, vol. 531, Part 2, pp. 370–388, 2015. doi: 10.1016/j.jhydrol.2015.08.045.

[15] S. Yatheendradas et al., "Understanding uncertainty in distributed flash flood forecasting for semiarid regions," Water Resources Research, vol. 44, no. 5, 2008. doi: 10.1029/2007WR005940.

[16] J. A. Falcone, D. M. Carlisle, D. M. Wolock, and M. R. Meador, "GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States," Ecology, vol. 91, no. 2, p. 621, 2010. doi: 10.1890/09-0889.1.

[17] J. A. Falcone, GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow. Reston, VA, USA: U.S. Geological Survey, 2011. doi: 10.3133/70046617.

[18] J. M. Johnson, *National Hydrologic Geospatial Fabric (Hydrofabric) for the Next Generation (NextGen) Hydrologic Modeling Framework*. HydroShare, 2022. [Online]. Available: http://www.hydroshare.org/resource/129787b468aa4d55ace7b124ed27dbde.

[19] J. Zhang et al., "Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities," Bulletin of the American Meteorological Society, vol. 97, no. 4, pp. 621–638, 2016. doi: 10.1175/BAMS-D-14-00174.1.

[20] W. Hanft, J. Zhang, and M. Simpson, "Dual-Pol VPR corrections for improved operational radar QPE in MRMS," Journal of Hydrometeorology, vol. 24, no. 3, pp. 353–371, 2023. doi: 10.1175/JHM-D-22-0010.1.

[21] A. P. Osborne, J. Zhang, M. J. Simpson, K. W. Howard, and S. B. Cocks, "Application of machine learning techniques to improve Multi-Radar Multi-Sensor (MRMS) precipitation estimates in the western United States," Artificial Intelligence for the Earth Systems, vol. 2, no. 2, Art. no. 220053, 2023. doi: 10.1175/AIES-D-22-0053.1.

[22] Q. Dong and F. Lu, "Performance assessment of hydrological models considering acceptable forecast error threshold," Water, vol. 7, no. 11, pp. 6173–6189, 2015. doi: 10.3390/w7116173.

[23] T. Bosshard, M. Carambia, K. Goergen, S. Kotlarski, P. Krahe, M. Zappa, and C. Schär, "Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections," Water Resources Research, vol. 49, no. 3, pp. 1523–1536, 2013. doi: 10.1029/2011WR011533.

[24] D. B. Wright, R. Mantilla, and C. D. Peters-Lidard, "A remote sensing-based tool for assessing rainfall-driven hazards," *Environmental Modelling & Software*, vol. 90, pp. 34–54, Apr. 2017, doi: 10.1016/j.envsoft.2016.12.006.

[25] D. Y. Hancock, J. Fischer, J. M. Lowe, W. Snapp-Childs, M. Pierce, S. Marru, J. E. Coulter, M. Vaughn, B. Beck, N. Merchant, E. Skidmore, and G. Jacobs, "Jetstream2: Accelerating cloud computing via Jetstream," in *Proc. Practice and Experience in Advanced Research Computing (PEARC '21)*, New York, NY, USA: Association for Computing Machinery, 2021, Art. no. 11, pp. 1–8, doi: 10.1145/3437359.3465565.

[26] T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth, and J. Towns, "ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support," in *Proc. Practice and Experience in Advanced Research Computing (PEARC '23): Computing for the Common Good*, Portland, OR, USA, Jul. 23–27, 2023. New York, NY, USA: Association for Computing Machinery, 2023, pp. 173–176, doi: 10.1145/3569951.3597559.

Chapter 4: Evaluating the Sensitivity of HAND Flood Inundation Predictions to River Slope Dynamics

Zih-Syun Chen[1], Reza Jamshidi[2], Sebastian Marshall[3], Pitamber Wagle[4], Anupal Baruah[5*], Sagy Cohen[6*]

[1]University of Houston; *zchen66@cougarnet.uh.edu*
[2]Northeastern University; *jamshidi.re@northeastern.edu*
[3]Johns Hopkins University; *smarsh46@jh.edu*
[4]Brigham Young University; *waglep@byu.edu*
[5]University of Alabama; *abaruah@ua.edu*
[6]University of Alabama; *sagy.cohen@ua.edu*
[*]Theme Leader

Abstract: The operational Flood Inundation Mapping (FIM) framework developed by NOAA's Office of Water Prediction generates flood inundation maps using Synthetic Rating Curves (SRCs) derived from Manning's equation, which depends on hydraulic parameters including channel roughness and slope. The current HAND-FIM framework utilizes static slopes from digital elevation models (DEMs) and the ICESat-2 River Surface Slope (IRIS) dataset, despite water surface slope varying with changing flow conditions and hydraulic processes. Surface Water and Ocean Topography (SWOT) mission provides repeated observations of water-surface elevation, enabling the estimation of time-varying water-surface slopes along river reaches. This study compares flood maps generated using existing static slope and three variants of dynamic slope from SWOT; median, 90th percentile, and slope corresponding to maximum water surface elevation. Simulations were performed for six areas in the United States using National Water Model streamflow data. The Big Sioux, Little Sioux, and Mississippi Rivers exhibited low variability in slope, resulting in simulated stages differing by only 1–4% across slope variants for the studied events. In contrast, the two Illinois River reaches and especially the Ohio River showed substantially greater sensitivity, with stage differences of 20–34% and up to 150%, respectively, consistent with backwater (slope falls with stage) influence. Despite these hydraulic differences, the flood extent changed little. For each slope product, the median CSI across the six cases ranged from 0.547 to 0.551, so an order-of-magnitude change in slope shifted the pooled flood-extent skill by only 0.004. Case-study-level CSI spreads ranged from 0.001 to 0.047 and were largest on the kinematic (slope rises with stage) Illinois reaches. Further, the consistent slope–discharge relationship between water surface slope and discharge indicates that dynamic, discharge-dependent slope estimation is feasible and could be incorporated into future operational flood inundation mapping frameworks.

1. Motivation

The operational Flood Inundation Mapping (FIM) framework of the NOAA OWP estimates water levels on each river reach using Synthetic Rating Curves (SRCs) developed with the Height Above the Nearest Drainage (HAND) methodology. The SRCs are governed/derived by calculating the discharge using Manning's equation at different synthetic stages. One of the key parameters impacting the quality of derived SRCs is the water surface slope. The hydrologic and hydraulic models commonly rely on river slopes derived from Digital Elevation Models (DEMs). The recent water surface slope products derived from satellite altimetry provide accurate estimates, and the average slopes from the Ice, Cloud, and land Elevation Satellite-2 (ICESat-2) River Surface Slope (IRIS) product are now being

incorporated into the latest version of FIM hydrofabric for wide rivers represented in the SWOT River Database (SWORD) [1]. However, slope is a dynamic component in river hydraulics, varying over both space and time. The recently launched Surface Water and Ocean Topography (SWOT) mission provides wide-swath altimetry of terrestrial water bodies, enabling extensive spatial and temporal coverage of global rivers [2]. In this context, this study addresses how to harness available water surface slope products to improve operational FIM.

2. Objective and Scope

This study is a proof of concept of the applicability of the time-varying slope in the current hydrofabric to improve the accuracy of HAND-FIM products, addressing the following three questions:

- How sensitive is the OWP HAND-FIM to satellite-derived slope from IRIS and SWOT?
- Does the current time series of SWOT slope records help improve FIM, and if so, in which locations and under what flow regimes?
- Using two case studies, can we differentiate regions based on the importance of the dynamic slope, and generalize the application?

3. Previous Studies

Measurement of river slopes has always been constrained by the need for accurate water surface elevation records at multiple points along the river centerline. Terrain-based products, including the Global River-Slope (GloRS), MERIT HYDRO, and HydroATLAS, use digital elevation models (DEMs) to estimate slope at a global scale based on the derived river lengths and elevation changes [3], [4], [5]. These datasets are foundational in large-scale hydrologic and hydraulic modeling, but are limited by the accuracy of derived slopes due to errors in DEMs [6]. The emergence of new techniques, such as the Global Navigation Satellite System (GNSS), has allowed for continuous measurements of river surfaces and estimation of slopes based on multiple cross sections along the river [7], [8]. These methods are either limited in scalability or do not provide simultaneous records along the stream for slope calculation, making estimates less reliable.

Launched in 2018, the ICESat-2 mission is the first altimetry satellite to measure water elevation based on six along-track beams, enabling slope calculation on a global scale. ICESat-2-derived slopes are standardized and aggregated into average slopes in the IRIS product, providing slopes on the SWORD network of reaches [9], [10]. Despite the high accuracy of derived water surface slopes from ICESat-2, its 91-day revisit time limits its ability to capture the temporal variability in river surface slopes [10]. SWOT's wide swath Ka-band radar enables simultaneous measurements of water surface elevation that are aggregated from PixelCloud into nodes 200 m apart along the river centerline, enabling reach-level slope calculation [2]. SWOT has a 21-day cycle with a median of two passes per cycle over the mid-latitudes, enabling the study of temporal dynamics in river systems [11], [12].

Slope is an important parameter in the development of the SRCs, which are governed by Manning's equation, and directly impacts the accuracy of derived flood inundation maps [14]. Chen et al. replaced the baseline slopes from the 3D Elevation Program (3DEP) with slopes from the IRIS dataset, which improved the inundation extents by 31±25% in terms of Critical Success Index across eight case studies [1], highlighting the significant influence of slope in the performance of operational flood maps.

The river surface slope is expected to remain constant for uniform flows; however, a time-varying slope is common across many rivers[12], [15]. Changes in riverbed slope, river confluences, and river–lake interactions can alter water surface profiles over time [15]. In large, low-gradient rivers, seasonal variations in flow regimes can induce temporal variability in water-surface slope, reflecting corresponding changes in stage and discharge conditions [16]. As demonstrated by Eggleston et al., there is no single-sided relationship between slope and discharge, with multiple patterns emerging across case studies [17].

4. Methodology

Figure 1 illustrates the three main components of this work: data selection, preprocessing, and analysis.

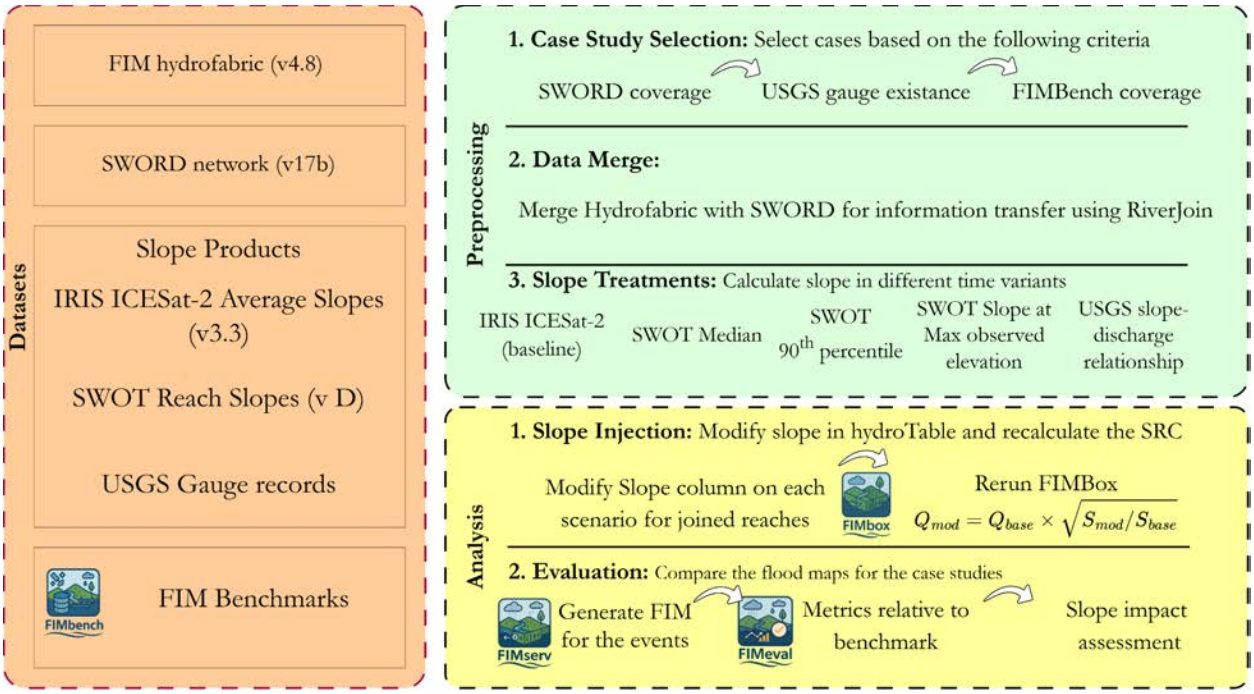


Figure 1. *Main steps in data acquisition, preprocessing, and flood inundation modification based on the resulting slope products across the case studies. The left panel portrays the datasets used, the top right panel shows the preprocessing steps, and the bottom right section demonstrates the core analysis framework.*

This study uses SWORD version 17b, as it is the operational version in the current SWOT data products, and the IRIS dataset [18]. We utilized version 4.8 of the FIM hydrofabric for modification and flood map generation inside the FIMbox framework. The temporally averaged slope from the latest version of the IRIS dataset (version 3.3) is used as the baseline. To control error propagation in derived slopes, we selected SWOT records based on the criteria mentioned in [19]. SWOT reach level records are accessed via the Hydrocron API [20]. The US Geological Survey (USGS) stream-gauge records were also utilized in locations with gauge pairs. Additionally, the reference flood maps used for evaluation were from FIMbench, restricted to aerial imagery, PlanetScope, and high-water mark-derived flood boundaries [21].

Six case studies were selected based on the coverage of the SWORD, the presence of at least one USGS gauge to enable continuous measurements and overcome SWOT data sparsity, and reference

flood map availability (**Figure 2)**. Two reaches on the Illinois River, along with reaches from the Little and Bis Sioux River, the Mississippi River, and the Ohio River, are selected based on the selection criteria. We assume that the selected case studies represent fluvial flooding events where variations in river slope directly influence the resulting flood inundation extents.

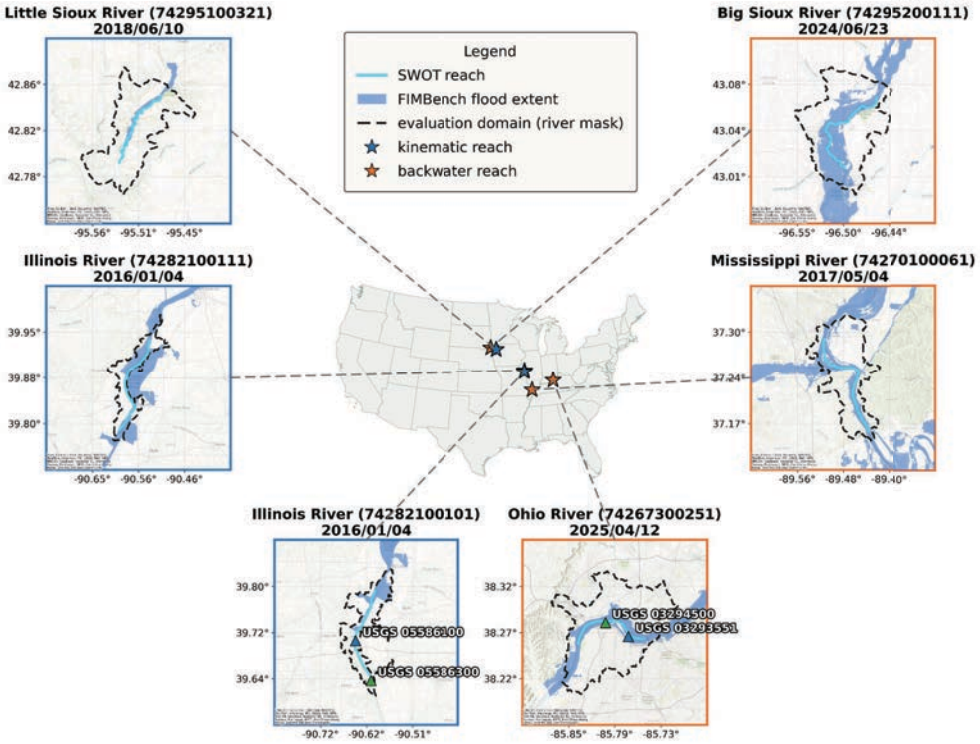


Figure 2. *Selected case studies (SWORD reaches) for slope impact assessment on HAND FIM.*

Since the SWOT data is built upon SWORD reaches and the HAND FIM framework uses its own hydrofabric, a crosswalk was necessary to transfer SWOT slopes to the HAND hydrofabric. This was done using the Riverjoin Python package [1]. We defined three slope treatments from the SWOT reliable record series. We computed the median as the expected slope, the 90th percentile as the slope when the flow is full in the channel and overtopping the bank, and kept the slope record corresponding to the maximum observed water surface elevation, representing the slope in the critical observed condition. We also generated the slope data from USGS paired gauges using long-term instantaneous records of gauge height at 15-min intervals. USGS gauge pairs lying within the SWORD reach were selected for this comparison. Among the six selected case studies, only two met this criterion. For these cases, we established the relationship between slope and discharge, and used this relationship to compute the synthetic rating curve.

To measure the impact of slope variability for each case study and derived slopes, the SRCs were regenerated utilizing the FIMserv Python package, an open-source framework for automated flood inundation mapping using geospatial data from OWP and river inputs [22]. All SRC calibration procedures available within the hydrofabric were disabled, including adjustments based on USGS streamflow gauges or other calibration sources, to isolate and quantify the impact of river slope modifications on the resulting flood inundation maps. The governing equation in the generation of SRCs is Manning's equation, in which slope contributes to discharge estimation through a square-root relationship.

$$Q = \frac{k}{n} A\, R^{\frac{2}{3}} S^{\frac{1}{2}} \quad (1)$$

Where Q is the discharge, k is the unit conversion factor, n is the roughness term, A represents cross-sectional area, R stands for the hydraulic radius, and S represents slope. Flood maps were developed by temporally matching each flood event with NWM Version 3.0. Events occurring before January 2023 used the Retrospective dataset, whereas events occurring during or after January 2023 used the Short-Range Forecast dataset. In both cases, the discharge is applied per feature identifier across the full contributing area, so each catchment is forced by its own flow rather than by a single reach value.

Following flood map generation, model performance was evaluated by comparing each simulated flood map with the available benchmark datasets. The spatial domain of evaluation for each case study was defined as the union of the catchments for which the slope was modified. Based on this pixel-level comparison, pixels are classified as True Positive (TP), True Negative (TN), False Positive (FP), or False Negative (FN). Metrics used in flood map comparison are summarized in **Table 1**.

Table 1. *Metrics for model performance based on the confusion matrix components.*

Metric	Definition	Equation
CSI (Critical Success Index)	Fraction of correct positives to all positives	$\frac{TP}{TP + FP + FN}$
F1 (F1-Score)	Harmonic mean of precision and recall	$\frac{2TP}{2TP + FP + FN}$
POD (Probability of Detection)	Fraction of actual positive events	$\frac{TP}{TP + FN}$
FAR (False Alarm Ratio)	Fraction of incorrect positive events	$\frac{FP}{TP + FP}$

5. Results

5.1. Slope choice reshapes rating curves but weakly affects flood-extent skill

The sensitivity of the synthetic rating curves to the choice of slope input was proportional to the spread among the input values (**Figure 3**). On the Big Sioux, Little Sioux, and Mississippi reaches (three of the six; Section 4.2), the slope inputs were nearly identical, and the resulting rating curves coincided. Stage at the event discharge varied by only 0.05 to 0.25 m (1 to 4%) on these reaches. On the two Illinois reaches, where the inputs differed moderately, the stage spread increased to 1.3 to 1.8 m (20 to 34%). The largest response occurred on the low-gradient, backwater-affected Ohio reach. There, the inputs spanned an approximately thirtyfold range (0.000036 to 0.0011 m/m), and the corresponding stage estimates ranged from 6.90 to 17.45 m, a spread of 10.6 m (per-input values in Supplement Section S1). Slope selection, therefore, exerted a direct control on the stage-discharge relationship, and its influence was greatest on low-gradient, hydraulically controlled reaches.

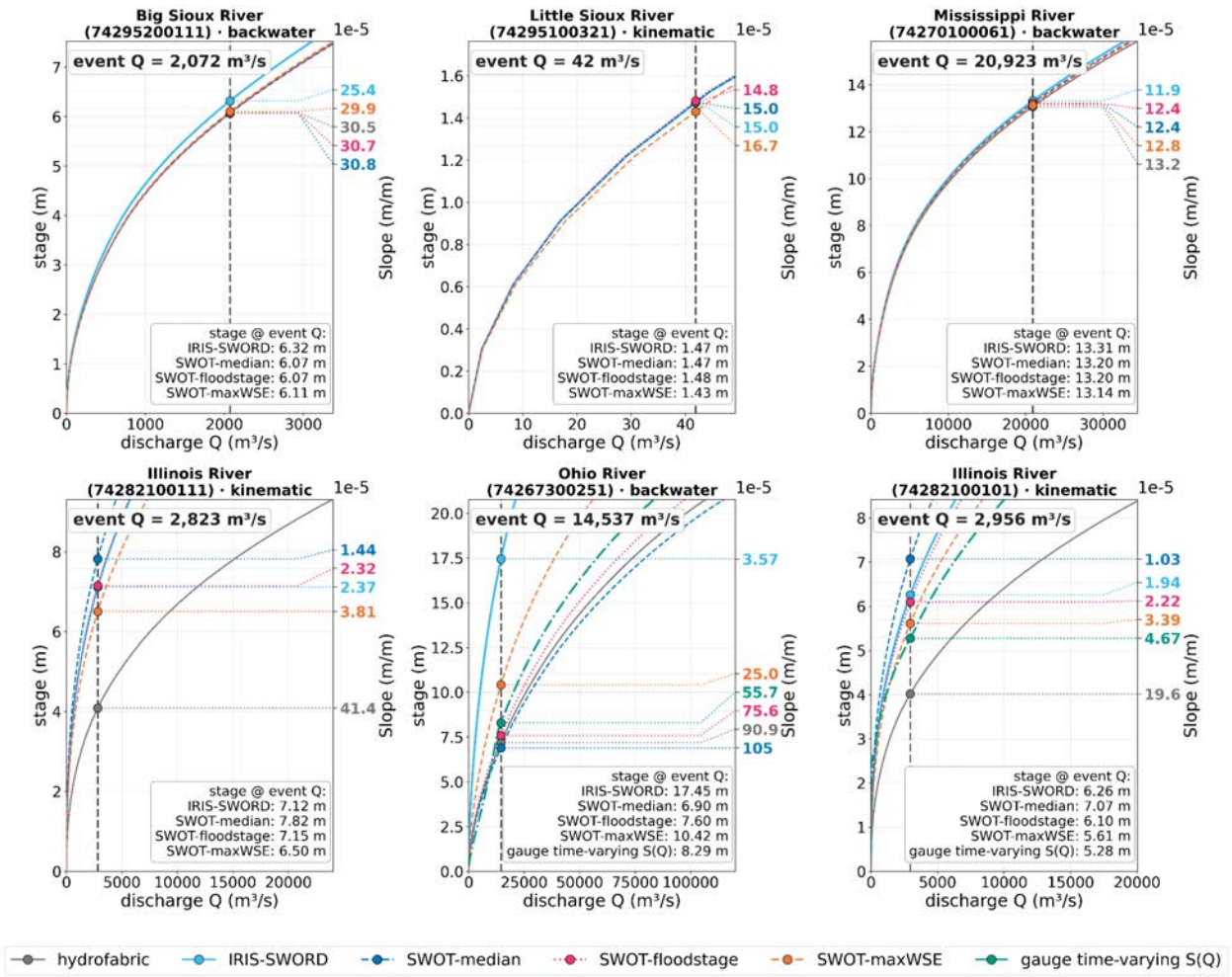


Figure 3. *Effects of slope products on synthetic rating curves. Rating curves are shown for six reaches under kinematic and backwater conditions. The vertical dashed line marks the event discharge, and the points indicate the corresponding stage estimated by each slope input, listed in the box the corresponding stage estimated by each slope treatment, listed in the box.*

These large hydraulic differences, however, produced only small changes in the mapped flood extent. Across the six case studies, the median CSI ranged from 0.547 for the SWOT products to 0.551 for IRIS-SWORD, a spread of only 0.004. The gauge-derived $S(Q)$, available on two case studies, reached a median CSI of 0.479. The contrast was sharpest on the Ohio reach, where stage estimates that differed by 10.6 m produced CSI values that differed by only 0.022 (Section 5.3). HAND-FIM was therefore strongly sensitive to slope in rating-curve construction but only weakly sensitive to slope in the mapped extent; the mechanisms behind this attenuation are examined in Supplement Section S2. Section 5.2 examines how slope varies with discharge on the study reaches, and Section 5.3 identifies where the choice of slope input measurably influenced the mapped extent.

5.2 Discharge-dependent slope reveals contrasting hydraulic regimes

Resolving how water-surface slope varies within a single flood is beyond the 21-day SWOT repeat cycle, so the discharge-dependent slope $S(Q)$ (Section 4.3) was reconstructed from continuous USGS records at pairs of gauges bracketing each reach (Section 4.1). Of the six case studies, only two gauge pairs produced a slope signal above the datum-related noise floor set by the gauge-datum and stage uncertainty over the pair separation (the 2σ band in **Figure 4**). These two reaches responded in

opposite directions. On the freely flowing Illinois River, slope increased with discharge as a power law (R^2 = 0.89; **Figure 4a**), consistent with kinematic or near-normal flow, in which the water surface steepens as discharge rises. On the backwater-affected Ohio River below McAlpine Dam, slope decreased as discharge increased (R^2 = 0.99; **Figure 4b**), consistent with a downstream control that flattens the water surface at high flow. Because the Ohio event discharge exceeded the fitted range, $S(Q)$ was held at the upper limit of the observed data rather than extrapolated (Supplement Section S3). The sign of the slope-discharge relationship thus distinguished the two regimes, with a positive relationship indicating freely flowing conditions and a negative one indicating backwater control. The diagnostic value of $S(Q)$ did not, however, extend to flood-extent skill (Section 5.3).

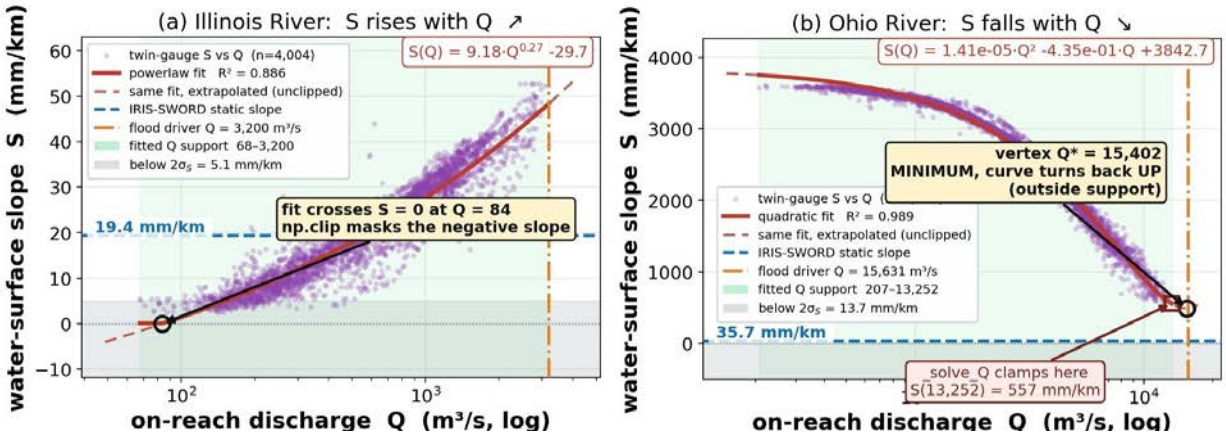


Figure 4. *Discharge-dependent water-surface slope:* ***(a)*** *Illinois River, where slope increases with discharge;* ***(b)*** *Ohio River, where slope decreases under backwater conditions. Shading shows the fitted discharge range, the blue dashed line shows the static slope, and the orange line shows the event discharge. On-reach discharge denotes the discharge assigned to the reach.*

Overall, **Figure 4** shows that the relationship between slope and discharge is useful for diagnosing the hydraulic regime (kinematic or backwater): increasing slope indicates a freely flowing response, whereas decreasing slope indicates backwater control. However, identifying the correct hydraulic behaviour did not consistently improve flood-extent skill relative to the static-slope products. Performance differences are examined in the following sections.

5.3. Flood-extent differences are spatially localized and vary more by reach than by slope product

For the Ohio River event, the five slope inputs produced broadly similar flood maps (**Figure 5**). True positives were concentrated along the main river corridor, false positives on the eastern floodplain, and false negatives along the channel margins and connected side channels. Differences among the maps occurred mainly along the edges of the inundated area. The metrics were consistent with this edge-confined pattern. IRIS-SWORD produced the highest CSI (0.354) but also the highest false alarm ratio (FAR, 0.577), whereas SWOT-maxWSE yielded a slightly lower CSI (0.342) and the lowest FAR (0.566; full per-product values in Supplement Section S4). Reducing false alarms therefore lessened overprediction beyond the observed extent, but at the cost of detection, a trade-off rather than a net gain in accuracy.

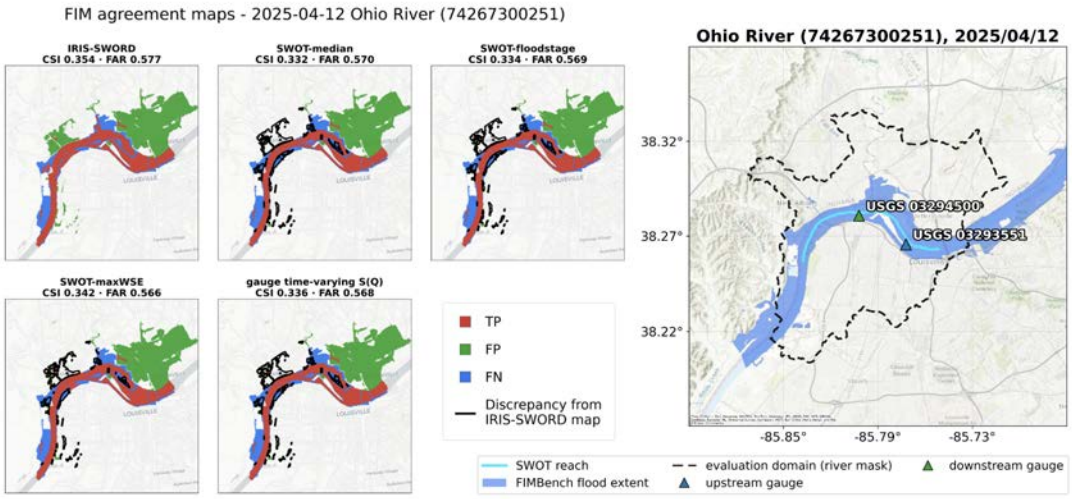


Figure 5. *Flood-inundation agreement for the 12 April 2025 Ohio River event: (first panel left-right) IRIS-SWORD; SWOT-median; SWOT-floodstage; SWOT-maxWSE; gauge time-varying S(Q). Red indicates true positives, green indicates false positives, blue indicates false negatives, and black outlines indicate differences from the IRIS-SWORD inundation map.*

Broadening the comparison to all six case studies (**Figure 6)** confirmed that the choice of slope input mattered least where the rating curves coincided. On the Big Sioux, Little Sioux, and Mississippi reaches, the flood maps were almost identical across inputs. Across the four static products, the per-case-study CSI spread ranged from 0.001 on the Mississippi to 0.047 on Illinois reach (74282100111), with the largest spreads on the two kinematic Illinois reaches (Supplement Section S4).

The clearest differences appeared on the Ohio and Illinois reaches. Currently within the scope of the cases evaluated, no single slope input ranked first on all four metrics (CSI, F1, probability of detection [POD], and FAR), so the preferred input depended jointly on the case study and the metric considered. The maps also differed in the direction of their error, from underprediction on the Sioux reaches to overprediction on the Ohio (full per-case metrics in Supplement Section S4). The two Sioux reaches underpredicted (0.56 to 0.62), Illinois reach 74282100111 was close to unbiased (0.94 to 1.15), and Illinois reach 74282100101 (1.25 to 1.43) and the Ohio reach (1.38 to 1.61) overpredicted.

Taken together, these results indicate that the variability among case studies exceeded the variability among slope inputs, and that the low skill scores did not stem from a single, shared bias. The influence of the slope input was instead specific to each reach, a pattern that favours reach-level diagnosis over a uniform slope correction.

6. Conclusion

This study addressed three questions: (1) how sensitive operational HAND-FIM is to alternative satellite-based slope products, (2) whether discharge-dependent slope improves flood mapping, and (3) where dynamic slope matters most hydraulically. Across the six SWOT-observed case studies, slope reshaped the SRCs in proportion to how much the slope inputs disagreed. Stage at the event discharge varied by 10.6 m on the Ohio and by 1.3 to 1.8 m (20 to 34%) on the two Illinois reaches, but by 0.25 m or less where the products agree. Flood-extent skill nonetheless varied little everywhere: the median CSI across the static slope inputs changed by only 0.004 (from 0.547 to 0.551), and several case studies showed nearly identical performance regardless of the assigned slope.

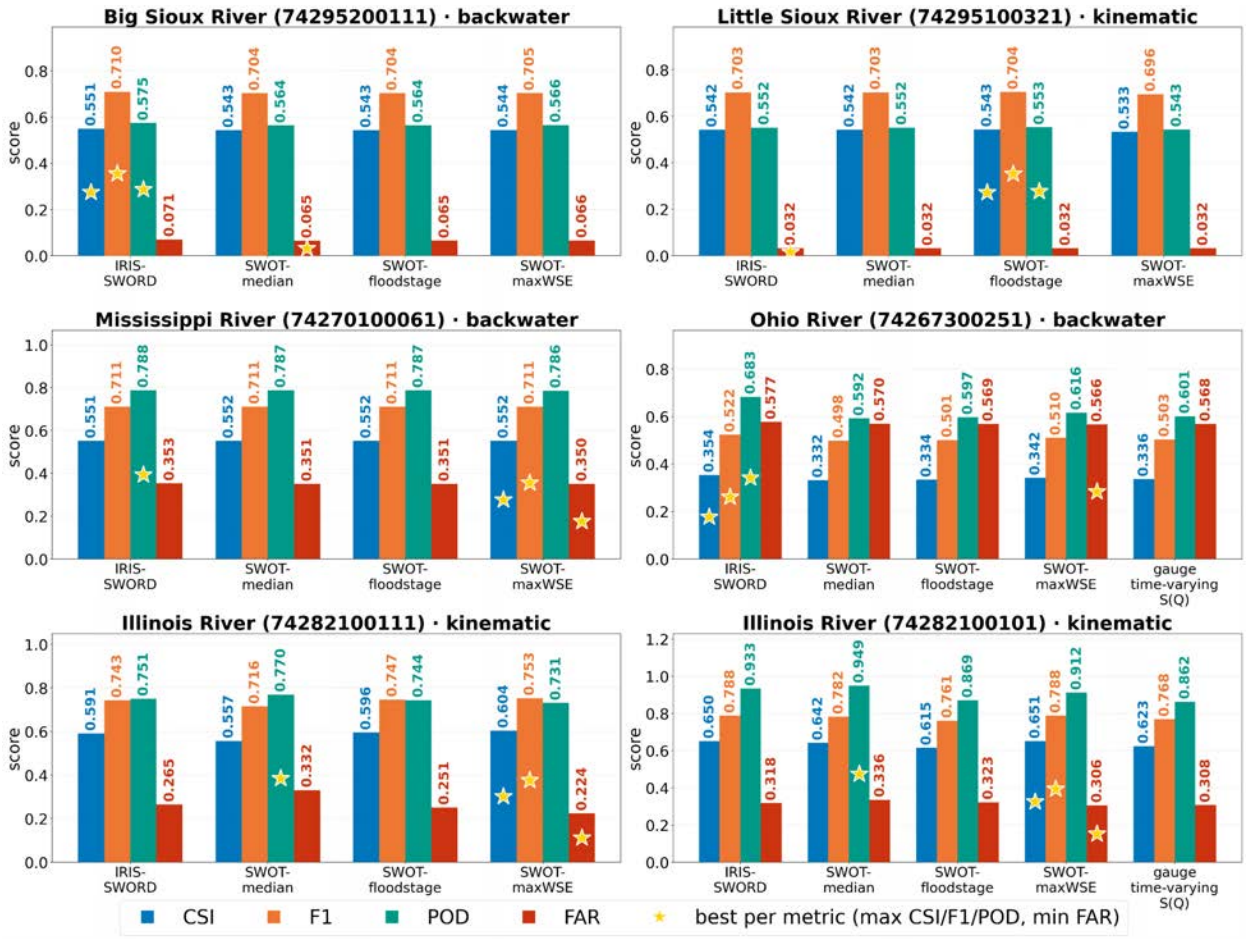


Figure 6. *Reach-level flood-inundation performance across slope treatments (left-right): Big Sioux River; Little Sioux River; Mississippi River; Ohio River; (Illinois River reach 74282100111; Illinois River reach 74282100101. Bars show CSI, F1, POD, and FAR, while stars identify the target-optimal value for each metric (CSI, F1, POD = 1; FAR = 0).*

The discharge-dependent slope treatment captured meaningful regime differences, with slope increasing with discharge on the freely flowing Illinois River and decreasing under backwater influence on the Ohio River. These relationships provide a physically interpretable diagnostic of the hydraulic regime, but the S(Q)-based maps did not consistently outperform static IRIS-SWORD or SWOT-derived slopes. Reductions in FAR often reflected more conservative flood extents that also came with more false negatives (missed observed flooding), so CSI, F1, and POD did not improve, underscoring that FAR must be interpreted together with CSI, F1, and POD. Thus, the primary value of S(Q) lies in diagnosing how a reach responds hydraulically, not in guaranteeing higher flood extent accuracy.

The conditions under which a dynamic slope is most relevant appear to be distinguished by hydraulic regime, although two $S(Q)$ case studies cannot establish a general pattern. HAND's inability to represent backwater is well documented: it assumes steady, uniform (normal) flow and cannot represent downstream hydraulic controls. These results add two things. First, the paired-gauge slope–discharge relationship provides an observation-based diagnostic that classifies the regime (Section 5.2). Second, even the regime-matched $S(Q)$ did not improve skill on the backwater Ohio case study, whereas on the kinematic Illinois reach 74282100111 the flood-condition slope products yielded the

best CSI and FAR. Regime screening is therefore proposed as a hypothesis for where dynamic slope is worth applying, to be tested across a larger sample, rather than as an operational rule.

These findings have practical implications for the operational use of HAND-FIM. The central decision for operational flood forecasters (NWM and OWP FIM users) is not which slope product is most accurate in isolation, but whether a reach is well represented by HAND's normal-flow assumption. For freely flowing reaches, satellite- or gauge-derived slope corrections can be justified where rating curve bias is a known error source. For dam-controlled, backwater, confluence-affected, or strongly nonuniform reaches, HAND outputs should be interpreted with caution and may need to be supplemented or replaced by models that explicitly represent backwater and wave dynamics, such as diffusive-wave or two-dimensional solvers. In this context, the slope–discharge relationship can serve as a screening tool to classify reaches, flag hydraulically unsuitable locations, and prioritize where additional modelling effort is warranted.

The conclusions remain constrained by the limited number of case studies and events, the availability of time-varying slope at only two paired-gauge locations, and compounded uncertainties in NWM discharge, DEM quality, channel geometry, and floodplain connectivity. Evaluation within the river mask (river-connected catchment) domain appropriately focuses on the area influenced by reach-based HAND modelling, but results should not be interpreted as characterizing off-channel, pluvial, tributary, or reservoir flooding. Future work should expand the set of reaches and events, develop quantitative, reproducible indicators (for example, the sign and magnitude of the slope-discharge relationship and its R^2) for distinguishing kinematic, transitional, and backwater regimes, and separating the contributions of slope uncertainty, discharge error, terrain, and connectivity to performance. Comparative studies with diffusive-wave and fully two-dimensional models will be needed to define when slope-adjusted HAND remains adequate and when alternative structures are required. Overall, the results suggest that slope is a direct control on synthetic rating curves, but only an indirect control on mapped inundation, and that hydraulic regime and model structure dominate flood-extent performance. Time-varying slope is therefore best used as a regime-informed diagnostic and selective correction, not as a universal replacement for static slope.

Acknowledgements: We express our sincere gratitude to our theme leads, Professor Sagy Cohen, and Dr. Anupal Baruah, for their invaluable guidance during this research. We are thankful to Dr. Yixian Chen for his continuous support. We thank Supath Dhital and Dipshika Devi for providing us with training on FIMbox and FIMeval. Sincere thanks to the Office of Water Prediction for sharing the ongoing efforts on improving FIM. We acknowledge the support of our academic advisors: Hyongki Lee (University of Houston), Edward Beighley (Northeastern University), Venkataraman Lakshmi (Johns Hopkins University), and Jim Nelson (Brigham Young University), whose support and encouragement facilitated this project and our professional development.

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA.

We acknowledge that ChatGPT (OpenAI, GPT-4.5) was used to correct the grammar and maintain the clarity. All the contents were thoroughly reviewed and edited by the authors thereafter, and assume full responsibility for the conclusions presented.

Supplementary Materials: All the processes and codes are available in the following GitHub repository: https://github.com/NWC-CUAHSI-Summer-Institute/TVSlope.

Supplementary Results & Discussion are available in the following document: https://www.hydroshare.org/resource/4b85cbe0a15a480fa13e4e7e61106fd7/.

References

[1] Y. Chen *et al.*, "Merging Remote Sensing Derived River Slope Datasets with High-Resolution Hydrofabrics for the United States," *Sci Data*, vol. 12, no. 1, p. 1657, Oct. 2025, doi: 10.1038/s41597-025-05941-6.

[2] L. Fu *et al.*, "The Surface Water and Ocean Topography Mission: A Breakthrough in Radar Remote Sensing of the Ocean and Land Surface Water," *Geophysical Research Letters*, vol. 51, no. 4, p. e2023GL107652, Feb. 2024, doi: 10.1029/2023GL107652.

[3] S. Cohen, T. Wan, M. T. Islam, and J. P. M. Syvitski, "Global river slope: A new geospatial dataset and global-scale analysis," *Journal of Hydrology*, vol. 563, pp. 1057–1067, Aug. 2018, doi: 10.1016/j.jhydrol.2018.06.066.

[4] D. Yamazaki, D. Ikeshima, J. Sosa, P. D. Bates, G. H. Allen, and T. M. Pavelsky, "MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset," *Water Resources Research*, vol. 55, no. 6, pp. 5053–5073, Jun. 2019, doi: 10.1029/2019WR024873.

[5] S. Linke *et al.*, "Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution," *Sci Data*, vol. 6, no. 1, p. 283, Dec. 2019, doi: 10.1038/s41597-019-0300-6.

[6] E. Rodríguez, C. S. Morris, and J. E. Belz, "A Global Assessment of the SRTM Performance," *photogramm eng remote sensing*, vol. 72, no. 3, pp. 249–260, Mar. 2006, doi: 10.14358/PERS.72.3.249.

[7] C. Schwatke, M. Halicki, and D. Scherer, "Generation of High-Resolution Water Surface Slopes From Multi-Mission Satellite Altimetry," *Water Resources Research*, vol. 60, no. 5, p. e2023WR034907, May 2024, doi: 10.1029/2023WR034907.

[8] L. H. Pitcher *et al.*, "AirSWOT InSAR Mapping of Surface Water Elevations and Hydraulic Gradients Across the Yukon Flats Basin, Alaska," *Water Resources Research*, vol. 55, no. 2, pp. 937–953, Feb. 2019, doi: 10.1029/2018WR023274.

[9] T. Markus *et al.*, "The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation," *Remote Sensing of Environment*, vol. 190, pp. 260–273, Mar. 2017, doi: 10.1016/j.rse.2016.12.029.

[10] D. Scherer, C. Schwatke, D. Dettmering, and F. Seitz, "ICESat-2 river surface slope (IRIS): A global reach-scale water surface slope dataset," *Sci Data*, vol. 10, no. 1, p. 359, Jun. 2023, doi: 10.1038/s41597-023-02215-x.

[11] C. Nickles, E. Beighley, Y. Zhao, M. Durand, C. David, and H. Lee, "How Does the Unique Space-Time Sampling of the SWOT Mission Influence River Discharge Series Characteristics?," *Geophysical Research Letters*, vol. 46, no. 14, pp. 8154–8161, Jul. 2019, doi: 10.1029/2019GL083886.

[12] Y. Chen, S. Cohen, S. Dhital, D. Devi, and A. Baruah, "Spatial and Temporal Variability in Global River Slopes from the Surface Water and Ocean Topography (SWOT) Satellite," Dec. 24, 2025. doi: 10.22541/essoar.176659889.96691731/v1.

[13] C. Fang *et al.*, "Satellite altimetry reveals intensifying global river water level variability," *Nat Commun*, vol. 17, no. 1, p. 958, Dec. 2025, doi: 10.1038/s41467-025-67682-9.

[14] F. Aristizabal *et al.*, "Extending Height Above Nearest Drainage to Model Multiple Fluvial Sources in Flood Inundation Mapping Applications for the U.S. National Water Model," *Water Resources Research*, vol. 59, no. 5, p. e2022WR032039, May 2023, doi: 10.1029/2022WR032039.

[15] P. Bauer-Gottwein, L. Christoffersen, A. Musaeus, M. C. Frías, and K. Nielsen, "Hydraulics of Time-Variable Water Surface Slope in Rivers Observed by Satellite Altimetry," *Remote Sensing*, vol. 16, no. 21, p. 4010, Oct. 2024, doi: 10.3390/rs16214010.

[16] C. M. Birkett, L. A. K. Mertes, T. Dunne, M. H. Costa, and M. J. Jasinski, "Surface water dynamics in the Amazon Basin: Application of satellite radar altimetry," *J. Geophys. Res.*, vol. 107, no. D20, Oct. 2002, doi: 10.1029/2001JD000609.

[17] J. Eggleston, C. Mason, D. Bjerklie, M. Durand, R. Dudley, and M. Harlan, "Siting Considerations for Satellite Observation of River Discharge," *Water Resources Research*, vol. 60, no. 6, p. e2023WR034583, Jun. 2024, doi: 10.1029/2023WR034583.

[18] E. H. Altenau, T. M. Pavelsky, M. T. Durand, X. Yang, R. P. D. M. Frasson, and L. Bendezu, "The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A Global River Network for Satellite Data Products," *Water Resources Research*, vol. 57, no. 7, p. e2021WR030054, Jul. 2021, doi: 10.1029/2021WR030054.

Chapter 5: Improving Flood Inundation Mapping through Stage-Dependent Manning's *n*

Ali Haider[1], Santiago Henao-Gomez[2], Armen Konialian[3], Alemayehu Dula Shanko[4], Anupal Baruah[5*], and Sagy Cohen[6*]

[1]The City College of New York, The City University of New York; *ahaider@ccny.cuny.edu*
[2]University of Iowa; *santiago-henaogomez@uiowa.edu*
[3]University of California, Los Angeles; *armenkonia@ucla.edu*
[4]Florida International University; *ashan036@fiu.edu*
[5]The University of Alabama; *abaruah@ua.edu*
[6]The University of Alabama; *sagy.cohen@ua.edu*
*Theme Leader

Abstract: The operational Flood Inundation Mapping (FIM) forecasting framework of NOAA's Office of Water Prediction relies on terrain-derived synthetic rating curves using Manning's equation to convert discharge to stage. Manning's roughness coefficient (n) is a major source of uncertainty. A constant value is assumed with respect to stage, with one value applied in the fixed channel and another once the stage reaches the floodplain. While computationally simple, this piecewise-constant scheme fails to capture the continuous change in roughness between channel and floodplain, or its spatial variability across terrain types and regions. This study aims to characterize the stage-dependent variability of Manning's roughness and test whether accounting for it improves FIM forecasts. Rating curves were calibrated by varying *n* across stages; at held-out rating-curve points, median absolute discharge error fell from 74.5% to 3.6%, improving at 98% of gauges. About four in five gauges showed stage-dependent n, most often increasing from channel to overbank, consistent with compound-channel hydraulics. Machine learning was then used to relate calibrated *n* to geomorphologic, geometric, and hydraulic variables and test regionalization to ungauged sites. Predictability was limited, likely due to small sample size, though adding Texas gauges (128 total) raised R^2 from below 0.1 to 0.19-0.31, highlighting a need for more data and CONUS-wide expansion. While gauge-based calibration substantially improved rating-curve accuracy, its benefit for predicting FIM was inconsistent, indicating a need for further study to better understand how calibrated roughness translates to FIM accuracy.

1. Motivation

NOAA Office of Water Prediction (OWP) produces operational Flood Inundation Mapping (FIM) using the Height Above Nearest Drainage (HAND) method [1], [2], which derives a reach-scale synthetic rating curve (SRC) from Manning's equation applied at each NHDPlus reach [3]. Discharge is delivered to each reach by T-Route, the National Water Model's channel routing engine [2]. In the operational configuration, Manning's *n* is represented using two constant values: 0.06 for the channel and 0.12 for the floodplain [2]. Where observed rating curves or benchmark inundation maps exist, OWP calibrates an adjusted roughness, but the adjusted value is held constant across all stages of a reach [2]. A machine learning-based extension of this correction has been proposed to cover unmonitored reaches [5]. In every configuration, *n* does not vary with stage. Once flow exceeds bankfull, the added conveyance is influenced by vegetation, obstructions, and floodplain microtopography rather than channel-bed friction alone [6]. Continental-scale analyses show that the

variability in roughness due to these above-bank features is substantial in both space and time [7], [8] and is not adequately represented by a single stage-invariant value.

A second limitation concerns the low-flow end of the rating curve. Digital Elevation Models (DEMs) record the water surface, not the channel bed, so the terrain-derived cross section omits the submerged channel, a limitation identified by OWP as a "known shortcoming and opportunity for improvement" as illustrated in **Figure 1** [2]. The synthetic and observed rating curves for a reach do not begin at the same discharge, and this offset is left uncorrected. Because *n* is calibrated by matching the synthetic rating curve to observations, the calibration process compensates for this discharge offset by adjusting n, rather than adjusting the missing channel geometry directly. As a result, the calibrated *n* reflects both channel roughness and the uncorrected geometric error, rather than roughness (friction) alone. This is a documented limitation of HAND applications [2].

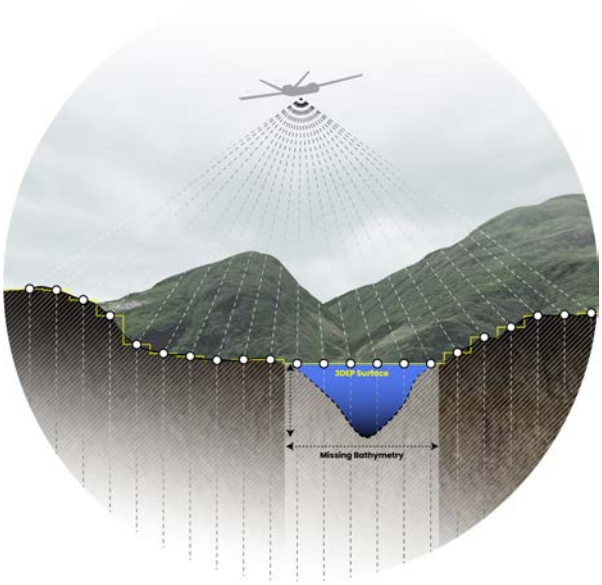


Figure 1. *DEM limitation: missing channel bathymetry* [2].

Existing corrections adjust the level of the rating curve without addressing its shape. Machine learning has been used to predict a per-reach adjustment factor across the conterminous United States [5] or rating curve power law parameters directly [9], and a separate correction replaces reach slope with ICESat-2 derived water-surface slopes, improving the critical success index by roughly 31% across eight flood events [10]. None of these make *n* a function of stage or correct the datum mismatch at the gauge scale. Johnson et al. [11] found that the coupled National Water Model (NWM) and HAND configuration underpredicts inundation even under perfect streamflow forcing, and traced part of that underprediction to the assumed Manning's n. To our knowledge, whether either error concentrates at the bankfull to overbank transition has not been tested in the literature reviewed here.

2. Objectives and Scope

The overarching objective of this project is to determine whether a spatially and stage-dependent representation of Manning's *n* improves the accuracy of OWP's operational HAND-FIM framework, relative to the fixed two-level (channel-floodplain) split or the single stage-invariant correction factor currently in use. Specifically, this study aims to:

1. Calibrate a stage-dependent Manning's *n* at USGS gauged reaches, using depth zones anchored at NWM recurrence flows.
2. Assess whether gauge-calibrated *n* values can be extended to ungauged reaches.
3. Quantify the effect of the calibrated and extended SRC on flood-map accuracy against independent FIM benchmarks.

Manning's *n* was calibrated at 102 of 164 candidate gauges across 23 HUC8 basins in North Carolina (NC) and South Carolina (SC). The remaining 62 gauges were excluded due to insufficient rating-curve data (46 gauges), recurrence flows falling outside the observed range of the rating curve (11 gauges), and duplicate HydroIDs (5 gauges). Six Texas Gulf HUC8 basins were later processed through the same pipeline, adding 26 calibrated gauges for a total of 128 across 29 HUC8 basins in two contrasting physiographic regions.

Extension of this methodology to ungauged reaches was tested in two ways: (1) machine learning regionalization from channel and catchment attributes, and (2) transfer of calibrated *n* along stream branches. Flood-map effects were evaluated against Hurricane Matthew (2016) benchmarks in the Neuse and Lumber basins and Hurricane Harvey (2017) benchmarks in the five Texas HUC8 basins. Future work will extend the approach to additional regions and ultimately the conterminous United States (CONUS).

3. Previous Studies

HAND was introduced by Rennó et al. [12]. HAND-FIM has a low computational cost, enabling operational flood mapping at continental scale. However, it requires stage-discharge relationships as an input. SRCs, analyzed and calibrated in this study, were developed to address this need. Their construction methodology was introduced by Zheng et al. [3], in which rating curves are derived from DEM-based channel hydraulic geometry and Manning's equation. Channel hydraulic geometry is estimated from cross sections extracted from the DEM at 0.3 m stage increments. This original methodology assumed a constant Manning's n. Subsequently, OWP implemented two-zones parameterization, assigning Manning's *n* values of 0.06 to the channel and 0.12 to the floodplain, as illustrated in **Figure 2**.

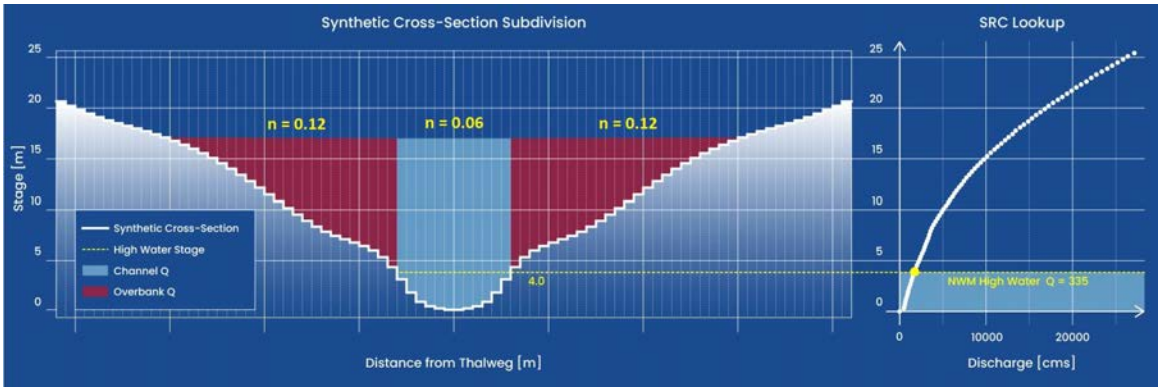


Figure 2. *Two-zones Manning's n parameterization of the DEM cross section [2].*

Several calibration and correction approaches have been proposed to improve the FIM performance derived from SRCs. In particular, OWP employs three strategies: calibration against USGS RCs, calibration against hydrodynamic models, and SRC adjustment factor (λ_{SRC}) [2]. Baruah et al. [5] used the λ_{SRC} available for 9000 gauges to train machine learning models and predict the factor over the

CONUS river network. The adjusted SRCs were evaluated using Hurricane Mathew, demonstrating improved overall FIM performance. Previous SRC calibration efforts have primarily focused on correcting the resulting rating curves through empirical adjustments or machine learning-based correction factors. In contrast, this study evaluates whether representing Manning's *n* as a stage-dependent parameter can improve SRC calibration.

Evidence that Manning's *n* varies with flow conditions has also emerged from recent studies. Mehedi et al. [13] used 200,000 surveys over 5000 sites from the HYDRoacoustic dataset in support of the Surface Water Oceanographic Topography satellite mission (HYDRoSWOT) to estimate the spatiotemporal variability of Manning's n. Manning's *n* was estimated from Manning's equation (1), using measured velocity and channel geometry. Their results showed that Manning's *n* tends to decrease and become more stable under high-flows conditions. These findings provide motivation for evaluating stage-dependent roughness in SRC calibration.

4. Methodology

The methodology follows the three objectives. First, the synthetic curves were calibrated by varying Manning's *n* at gauged sites across discrete stage intervals. Here the spatial and stage-dependent variability of Manning's n is evaluated. Second, the transferability of calibrated Manning's *n* values from gauged to ungauged streams was evaluated using machine learning models and geometric and hydraulic characteristics derived from the gauged sites. Third, HAND-FIM accuracy was assessed when these stage-dependent Manning's *n* estimates are incorporated into the synthetic rating curves (**Figure 3**). Each of these steps are described in further detail in Section 4.2.

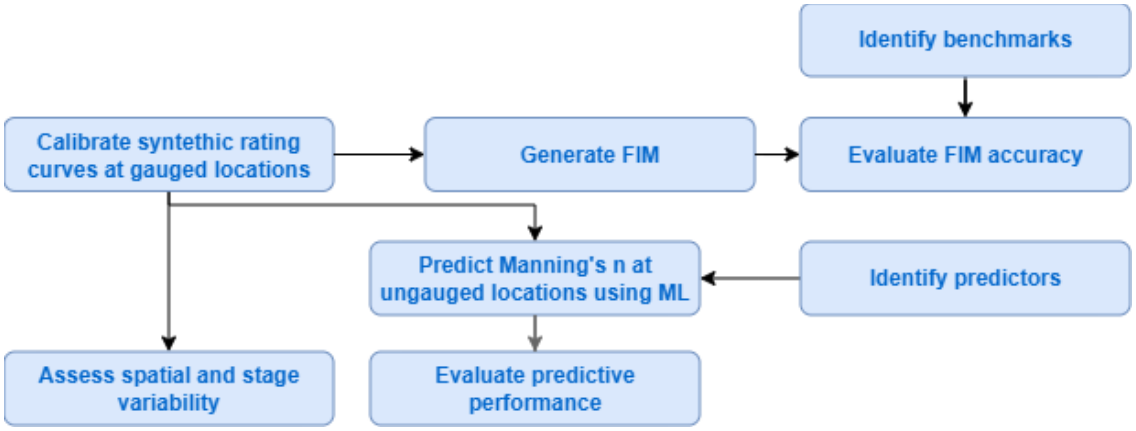


Figure 3. *High-level overview of the methodological framework used in this study.*

4.1 Study area

The Carolinas and Texas were selected due to the availability of USGS gauges with published rating curves and independent flood inundation benchmarks in those regions. A total of 102 gauges across 23 HUC8 basins in coastal North Carolina and South Carolina and 26 gauges across six Texas Gulf HUC8 basins were used to calibrate the synthetic rating curves and train the machine learning models (**Figure 4**). The two regions also provide contrasting physiographic and hydrologic settings for evaluating transferability. FIM performance is benchmarked against Hurricane Matthew (2016) in North Carolina [5], [8], [11], and Hurricane Harvey (2017) in Texas [5], widely used in the HAND-FIM literature.

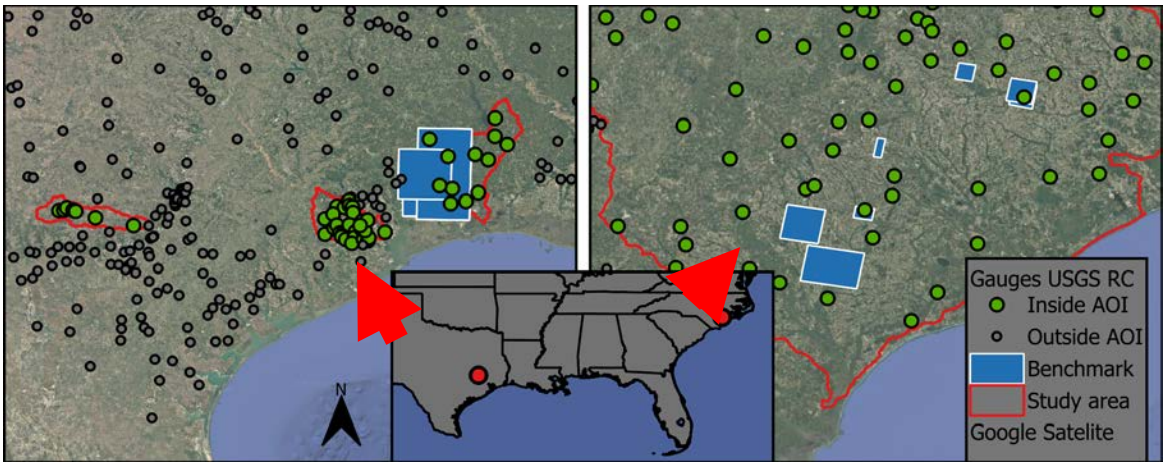


Figure 4. *Location of the study areas and used benchmark datasets. Left: Texas basins; right: North Carolina and South Carolina basins.*

4.2 Stage-dependent Manning's n calibration

The calibration process was performed using a modified version of FIMbox, a Python implementation of NOAA OWP's operational HAND-FIM framework (inundation-mapping v4.9.x) [14]. HAND-FIM computes wetted cross-sectional area A, hydraulic radius R, and bed slope S from the DEM, at 0.3048 m stage increments, then relates stage to discharge through Manning's equation [3]:

$$Q(h) = \frac{1}{n} A(h) R(h)^{\frac{2}{3}} S^{\frac{1}{2}} \tag{1}$$

where $Q(h)$ is the discharge (m^3/s) at stage h (m), n is Manning's roughness coefficient, $A(h)$ is the wetted cross-sectional area (m^2), $R(h)$ is the hydraulic radius (m), and S is the dimensionless bed slope. A, R, and S are computed from the terrain and are therefore fixed once the DEM is processed; n is the only quantity the data do not determine, and the operational OWP HAND-FIM framework, which FIMbox reproduces, holds it constant across all stages of a reach [2]. This calibration resolves n instead as a piecewise function of stage at each gauge (**Figure 5**).

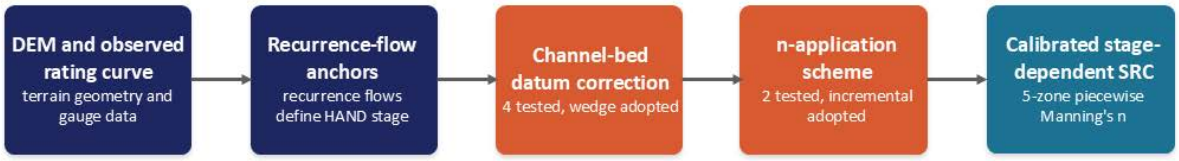


Figure 5. *Stage-dependent Manning's n calibration pipeline.*

4.2.1 Recurrence-zone anchor

At each gauge, the 2-, 5-, 10-, 25-, and 50-year NWM recurrence discharges are identified through the observed USGS rating curve to obtain an elevation, converted to HAND stage. These stage-discharge pairs serve as anchors, points at which the depth-discharge relationship is known from gauge observations and through which the calibrated curve is constrained to pass. The anchors bound five depth zones within which n is back-calculated independently and held piecewise constant. Not every gauge has all five anchors, because the observed rating curve must extend to a given recurrence stage

for that anchor to exist; this heterogeneity is carried into the reporting rather than averaged away. An example for gauge 02111180 in NC is shown in **Figure 6**.

4.2.2 Correcting the channel-bed datum

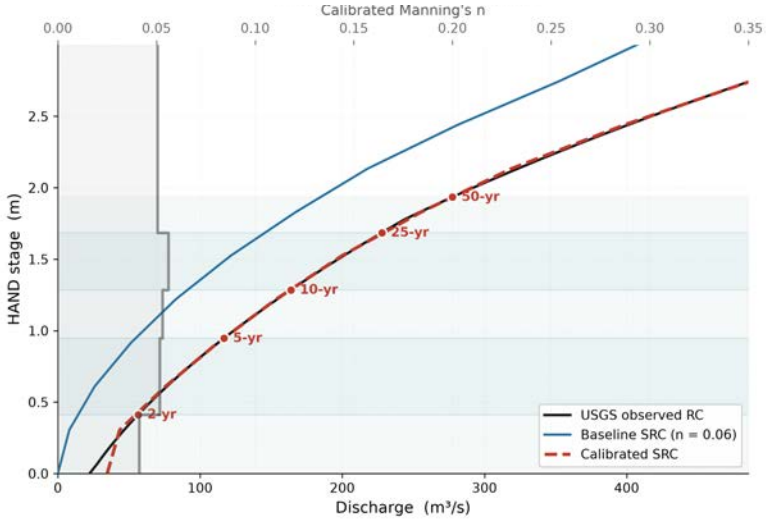


Figure 6. *Example of the recurrence-zone anchor scheme at gauge 02111180 (HUC 03040101). The observed USGS rating curve (black) is shown with the uncalibrated synthetic rating curve at n = 0.06 (blue) and the calibrated synthetic rating curve (red dashed). Red points mark the recurrence-flow anchors, and the gray line, read against the top axis, shows the resulting piecewise Manning's n. The datum gap at this gauge is δ = 0.77 m.*

Since, DEMs resolve the water surface rather than the channel bed, a HAND stage, depth of water measured from the thalweg, of 0 m already includes real baseflow, whereas the uncorrected synthetic curve reports a discharge of 0 m^3/s. Calibration against this curve systematically underestimates n, because the understated area and hydraulic radius leave a reduction in *n* as the only means of matching observed discharge, and the calibrated value then compensates for missing geometry rather than representing friction. Four candidate corrections were implemented and compared: a bathymetry wedge of fixed width, a datum shift, a baseflow offset as shown in Garousi-Nejad et al. [8], and a bathymetry wedge whose width is solved from Manning continuity so that the wedge carries the observed baseflow by construction,

$$Q(h) = \frac{1}{n} A(h) R(h)^{\frac{2}{3}} S^{\frac{1}{2}} \quad (2)$$

where Q_0 is the observed baseflow (m^3/s) read from the rating curve at the DEM water-surface elevation and δ is the datum gap (m). The comparison used 94 gauges and 384 anchors, slightly fewer than the production run because all four methods were required to fit at each gauge simultaneously. All four produced comparable held-out error, about 11 to 15%, since *n* absorbs whatever the geometry correction misses. Therefore, accuracy alone cannot discriminate between them. The continuity-solved wedge was adopted because it keeps *n* physically interpretable. Among the four methods it shows the lowest saturation at the physical bounds of n, its median calibrated value falls at the default of 0.06 without being forced there, and 76% of its values lie within the natural-channel range of Chow [6]. In addition, the modeled and observed curves begin at the same discharge by construction, rather than by coincidence.

The datum gap δ is estimated per gauge by fitting the low-flow limb of the rating curve to the standard USGS power law [15], [16]:

$$Q = C(elev - h_0)^b \quad (3)$$

where Q is discharge (m^3/s), elev is the water surface elevation (m), h_0 is the fitted elevation of effectively zero discharge (m), b is a dimensionless fitted exponent reflecting the hydraulic control, and C is a fitted scale coefficient dependent on b. The datum gap δ (m) is the difference between the DEM zero-flow elevation and h_0. A fit is accepted only if it passes three checks: b must lie between 1.0 and 5.0, the range of hydraulically plausible control exponents; R^2 must be at least 0.95, confirming that the power law actually describes the low-flow limb; and δ must fall between 0.05 and 15.0 m, since smaller gaps are indistinguishable from elevation noise and larger ones indicate a datum or crosswalk error rather than missing bathymetry. Gauges failing any check fall back automatically to the baseflow-offset correction [8].

4.2.3 Applying n across the recurrence-flow depth zones

Calibration produces one n value per recurrence-flow depth zone, and how that value is applied to the cross section determines whether the resulting curve is physically valid. Applying each zone's n to the entire wetted cross section recomputes the discharge of all previously flowing water under the new roughness at every zone boundary. Where adjacent zones differ substantially, as when flow first engages a rough floodplain, this recomputation can reduce total discharge as stage rises, which is hydraulically inadmissible and corrupts the flow-to-stage lookup used in FIM generation, since one discharge can then correspond to multiple stages.

The incremental scheme avoids this by leaving previously computed flow untouched: each zone's n applies only to the depth band newly wetted between consecutive anchors, and its contribution is added to the discharge already accumulated below. The n for zone i is solved from the conveyance and discharge added between the bracketing anchors:

$$n_i = \frac{\sqrt{S}(K_i - K_{i-1})}{Q_i - Q_{achieved,i-1}}, where\ K = AR^{2/3} \quad (4)$$

where K_i is conveyance at anchor i, Q_i is the recurrence discharge anchoring zone i, and $Q_{achieved,i-1}$ is the discharge accumulated through the previous zone under this same scheme. Because the rising stage can only add conveyance, the curve is continuous and monotone by construction.

In a direct comparison on the same 94 gauges, the incremental scheme reduced held-out error nearly fourfold relative to whole-section application, from 13.1 to 3.6 %, with bias falling from −8.0 to −1.1 %, because each depth band is constrained by both bracketing anchors rather than only its upper one. The cost is a rise in bound saturation, from 9.5 to 11.6 % of anchors at the upper bound and from 2.6 to 4.6 % at the lower, since solving n from the flow increment between anchors is noisier than solving it from the full anchor flow. This trade was accepted given the accuracy gain. Both roughness definitions are retained as parallel outputs, the incremental n as the calibrated variable and an independently recomputed whole-section n solely for comparability with OWP's 0.06/0.12 convention.

4.3 Generation of SRC at ungauged location using machine learning

Following the zone-wise calibration of Manning's *n* at gauged reaches across North Carolina, South Carolina, and Texas, machine learning models were developed to evaluate whether Manning's *n* can be predicted at ungauged reaches from catchment and basin characteristics in gauged locations. Geomorphologic, geometric, and hydraulic variables, consistent with those analyzed by Baruah et al. [9], were used as predictor variables while the calibrated Manning's *n* values corresponding to the recurrence interval stage segments of 2-, 5-, 10-, 25-, and 50-year flows, together with the baseflow parameter (h_0), were used as the target variables. Four machine learning algorithms—XGBoost, Ridge regression, Lasso regression, and Random Forest—were evaluated as an initial prediction benchmark. These algorithms were selected because they are widely used in hydraulic prediction applications. Rather than training a single multi-output model, separate models were developed for each target variable such that feature importance is evaluated separately for different Manning's *n* zones. Predictors include: catchment average silt %, and seasonal vegetation indices (EVI) [17]; average catchment slope %, stream order, total drainage area, and average discharge [18]; average precipitation, and mean hydraulic conductivity [19]; median bed particle size [20]; soil characteristics [21]; and DEM-derived hydraulic geometry.

4.4 Flood inundation mapping and evaluation

FIMbox was used to generate flood inundation maps. FIMbench and FIMeval, both open-source tools within the same ecosystem [14], were used to retrieve benchmark inundation maps and compute accuracy metrics relative to those benchmarks, respectively. The SRCs derived from the preceding calibration and machine learning steps were used as inputs to generate flood inundation maps. The generated SRCs were evaluated against the baseline FIM to quantify whether performance improves. This approach used calibrated Manning's *n* values derived at gauged reaches, transferred to the branch on which each reach is located. Furthermore, if a branch contains more than one USGS gauge, the *n* value from the nearest gauge is adopted. This configuration was evaluated at branches with available flood benchmarks to assess how well calibrated Manning's *n* performs, rather than generalizing the approach to all ungauged reaches. The primary benchmarks available for the study area are High Water Marks (HWMs) and a Post-Storm Survey (PSS), both derived from ground observations collected during and after the flood event. Because HWMs represent the maximum water surface elevation reached at each surveyed point rather than a snapshot at one moment in time, they capture multiple peak conditions across the event, not the inundation extent at any single instant. Comparing these peak flow conditions directly against a FIMbox map generated for one fixed timestamp would be misleading, since individual reaches across a river network do not peak simultaneously. To address this, inundation was generated independently for each reach at its own peak discharge, and the reach-level results were then mosaicked into a single composite map for each watershed, making the result directly comparable to the benchmark's peak-extent inundation.

This approach was validated by comparing the modeled FIM with the benchmarks using multiple complementary performance metrics, as no single metric can fully characterize flood inundation mapping (FIM) accuracy. The evaluation included the Critical Success Index (CSI), Probability of Detection (POD), False Positive Rate (FPR), and F1 Score. These metrics were computed from the confusion matrix, which consists of True Positives (TP; predicted wet and observed wet), False Positives (FP; predicted wet and observed dry), False Negatives (FN; predicted dry and observed wet), and True Negatives (TN; predicted dry and observed dry). CSI, POD, and F1 range from 0 to 1, where higher values indicate better agreement between predicted and observed inundation (1 being a perfect match). FPR also ranges from 0 to 1, but in the opposite sense: lower values are better, with 0

indicating no dry areas were falsely predicted as wet. The mathematical expressions for these performance metrics are provided below.

$$CSI = \frac{TP}{(TP + FP + FN)} \tag{5}$$

$$POD = \frac{TP}{(TP + FN)} \tag{6}$$

$$FPR = \frac{FP}{(TN + FP)} \tag{7}$$

$$F1 = \frac{2TP}{(2TP + FP + FN)} \tag{8}$$

5. Results

5.1 Stage-dependent Manning's n calibration

Of the 102 calibrated gauges, 99 through the continuity-solved wedge correction and 3 through the automatic baseflow-offset fallback where the datum fit failed quality control. Calibration at these gauges produced 383 zone-level *n* values, one per recurrence-flow anchor. The median calibrated *n* is 0.067. 72.3% of values fall within the natural-channel range of 0.02 to 0.20 [6], with 11.5% pinned at the upper physical bound of 0.30 and 4.7% at the lower bound of 0.01.

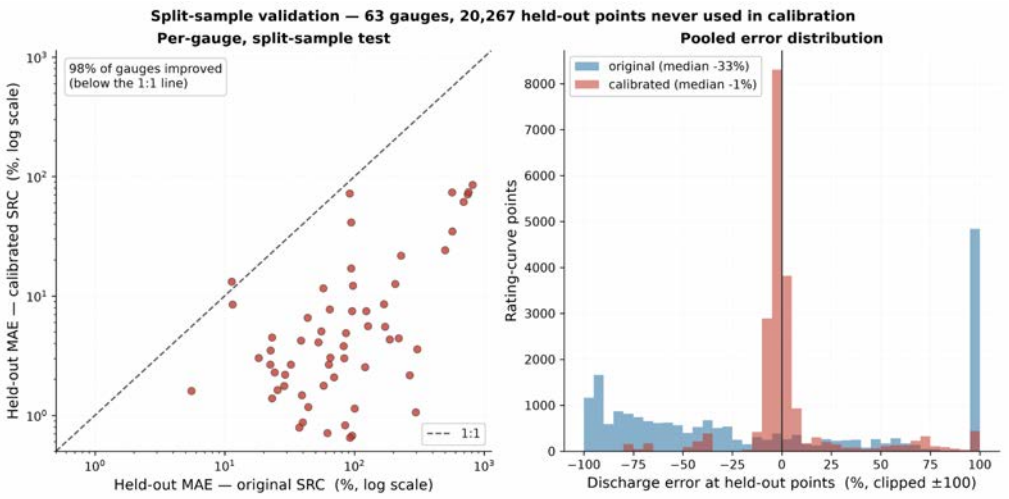


Figure 7. *Split-sample held-out validation across 63 gauges, 20,267 observations never used in calibration. Left: per-gauge held-out median absolute error, original versus calibrated synthetic rating curve, on log-log axes; 98% of gauges fall below the 1:1 line. Right: pooled distribution of signed discharge error at held-out points, clipped at ±100%.*

Calibration accuracy was evaluated using a held-out validation approach. Because the anchors serve as fitting targets, agreement at those points does not constitute independent evidence of accuracy. The evaluation therefore used the USGS rating-curve points falling strictly between the calibration anchors, which were excluded from fitting. Across the 63 gauges with such observations, 20,267 points were compared against the original and calibrated synthetic rating curves. Median absolute error (MAE) was between 3.6 and 74.5%, median signed error (MSE) moved from −32.6 to −1.0%, and 98% of

gauges improved (**Figure 7**). Because none of these points informed the calibration, this measures generalization within each gauge's observed range rather than fit to the anchors. Whether the calibrated stage dependence is physical, rather than a measurement artifact, was assessed before reporting its magnitude.

Before interpreting the calibrated stage dependence, each gauge was screened for whether its *n* profile can be trusted. Gauges were classified as clean, suspect, or reject using two criteria: whether any of the gauge's calibrated *n* values sit at the imposed physical bounds of 0.01 or 0.30, which indicates the calibration was constrained rather than converged, and the quality of the datum-gap fit. This gave 67 clean, 32 suspect, and 3 reject gauges. The rejects have an irrecoverable datum mismatch; at gauge 02088383, for example, the observed rating curve begins several meters above HAND stage zero, a gap no roughness value can close. All shape statistics below use the clean gauges only.

Each clean gauge's sequence of zone *n* values traces a profile against stage, classified into four shapes: increasing, decreasing, hump (rising into the overbank zones and then falling at the deepest zone), or mixed. Detecting a hump requires at least three anchors, since a reversal needs a point on each side of the peak, so gauges with short observed rating curves and few anchors can only ever register as increasing or decreasing, whatever their true shape. The apparent shape distribution therefore depends on how much of each gauge's stage range was observed. This shows directly in the data: restricting the census to gauges with at least two, three, four, and five anchors, the hump fraction rises steadily from 32 to 53% while the increasing fraction falls from 40 to 16%. The likely reading is that many gauges classified as increasing are humps whose reversal lies beyond their observed range, and the hump fractions reported here are consequently lower bounds.

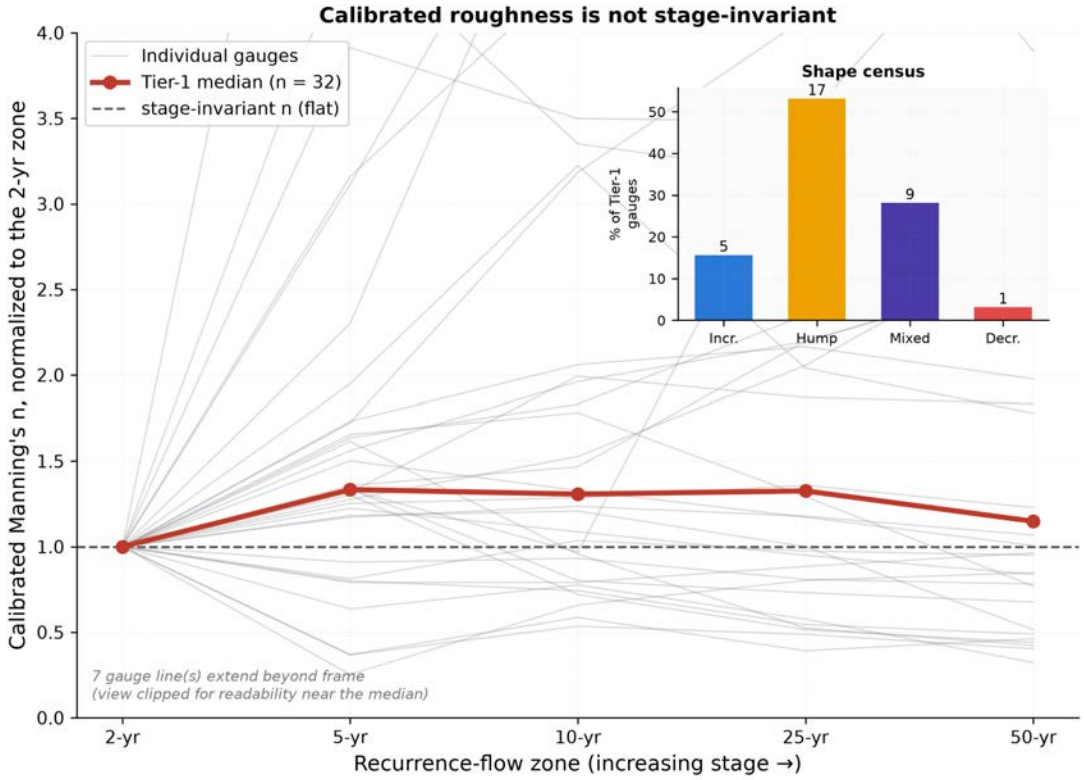


Figure 8. *Stage dependence of calibrated Manning's n among the 32 gauges with a complete five-anchor profile, normalized to each gauge's own 2-year value. Individual gauge trajectories in gray behind the median. Inset: shape census of the same 32 gauges.*

Magnitude is reported at two tiers. Among the 32 gauges with a complete five-anchor profile, the median *n* normalized to each gauge's own 2-year value runs 1.00, 1.33, 1.31, 1.33, and 1.15 across the five zones, a rise of roughly one third into the overbank zones that eases at the 50-year zone (**Figure 8**); only 22% of these gauges are within 25% of flat. Among the 63 clean gauges with at least two anchors, stage dependence was summarized as the ratio of the calibrated *n* at each gauge's highest available anchor to that at its lowest. The median ratio is 1.78, meaning roughness nearly doubles across the observed stage range at the typical gauge. The ratio exceeds 1 at 67% of gauges and falls below 1 at 33%, and only 21% of gauges lie within the range of 0.75 to 1.25 where a single constant *n* would be adequate. Because rating-curve coverage varies, the highest anchor differs among gauges, so this ratio is a per-gauge summary rather than a uniform 50-year to 2-year comparison. The 32 five-anchor gauges are a subset of the 63, so the second tier shows that the conclusion survives relaxing the anchor-count requirement rather than providing independent replication. In both tiers, roughly four in five gauges cannot be adequately represented by a single stage-invariant n, and the dominant direction is an increase from channel to overbank, the floodplain-roughening signature that compound-channel hydraulics predicts [6]. The partial reduction in *n* value at the deepest zone is directionally consistent with the decrease of roughness under high flows reported by Mehedi et al. [13], although the dominant increasing pattern found here differs from their overall finding.

To evaluate how the calibrated Manning's *n* varies across recurrence intervals, the 2-yr through 50-yr values were compared for each gauge, and gauges with fewer than four available values were excluded. The remaining sequences were classified as increasing, decreasing, or no trend based on monotonicity (ties counted as breaks), allowing a single deviating value to be dropped if doing so restored a consistent trend (**Figure 9**).

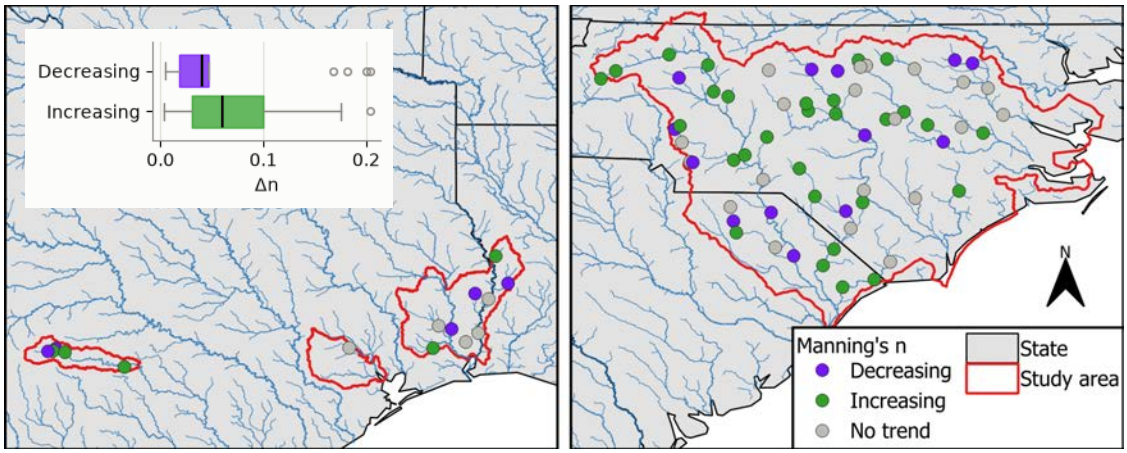


Figure 9. *Spatial distribution of Manning's n trend classifications.*

5.2 Generation of SRC at ungauged location using machine learning

The five calibrated zone Manning's *n* values and h_0 served as target variables. Trained on the 102 North Carolina and South Carolina gauges, with effective sample sizes of 61 to 99 depending on the target variable (Manning's *n* zone and h_0), all models performed poorly for the *n* zone targets, with R^2 below 0.1, while h_0 was consistently predicted well (R^2 above 0.8). Incorporating the 26 Texas gauges raised the sample to 128 and improved the Manning's *n* models to R^2 of 0.19 to 0.31 (cross-validated) after hyperparameter tuning and feature engineering. The improvement with added gauges suggests the models are data-limited at this sample size, and further gains are expected as additional regions are calibrated.

5.3 Flood inundation mapping and evaluation

Flood inundation maps generated using the technique of transferring calibrated Manning's *n* values to the nearest branches, which we named the 'branch-level n-transfer calibration method,' were evaluated against all available benchmark datasets in the study area. Benchmarks were available for the Neuse basin, which covers HUC03020201 and HUC03020202, and the Lumber basin (HUC03040203) in North Carolina and five HUCs in Texas. The benchmarks are for Hurricane Matthew in North Carolina and Hurricane Harvey in Texas. These datasets are heterogeneous (survey-based versus single-timestamp), and each covers only a small fraction of the HUC's area.

Calibration improved CSI in the Neuse basin across four evaluation benchmarks while producing the opposite effect in the other three benchmarks for the Lumber basin (**Table 1**). This divergence was anticipated by a rating-curve-level transfer test conducted before generating flood maps. On branches with two or more calibrated gauges, each gauge's calibrated *n* values were transferred to the other gauge and scored against that gauge's held-out rating-curve observations. The donor gauges on the Lumber branch produced errors between 84 and 97%, while the Neuse donors, serving evaluated benchmark reaches, showed errors of 2 to 3%. The case for the Neuse basin is illustrated in **Figure 10,** where the event-max benchmark (code-named hwm2nc), covering the Hurricane Matthew event, is compared against the OWP (baseline) and n-transferred configurations.

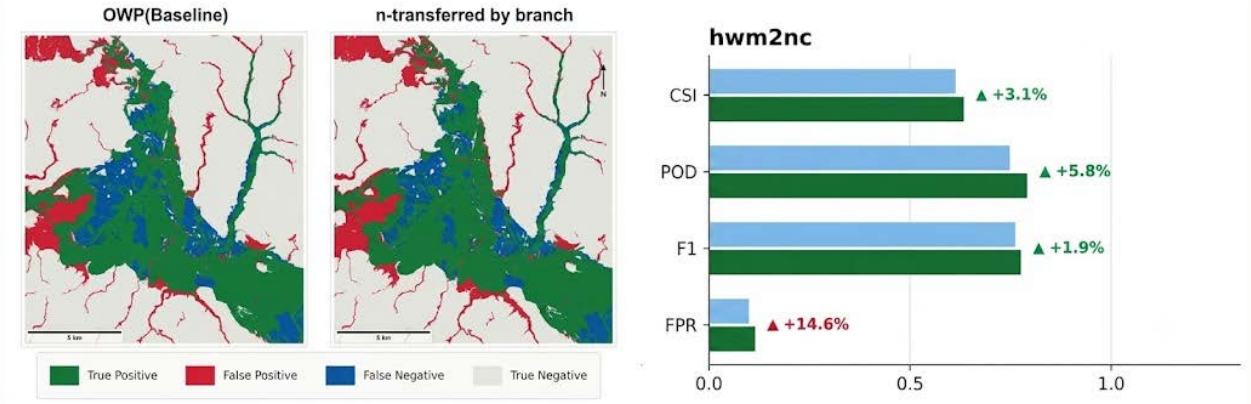


Figure 10. *Flood inundation mapping (left) and evaluation metrics for the Neuse basins (right). The OWP (baseline) evaluation metrics shown in cyan and branch-level n-transfer calibration shown in green are evaluated against the event max benchmark of Hurricane Matthew (hwm2nc). Percentage labels show the change from baseline to calibrated, colored green for improvement and red for degradation (lower FPR counts as improvement).*

In Texas, HUC12010005 degraded slightly across CSI and POD, while the downstream HUCs, HUC12020003, HUC12020006, and HUC12020007, saw CSI and POD improve at the cost of higher false-positive rates; HUC12040102, validated against an independent aerial-imagery-derived map, showed the same pattern.

Table 1. *CSI performance under branch-level n-transfer calibration versus baseline by HUCS and benchmark for North Carolina and Texas.*

HUC ID	Benchmark	CSI			POD			FPR		
		Baseline	Calibrated	ΔCSI	Baseline	Calibrated	ΔPOD	Baseline	Calibrated	ΔFPR
North Carolina										
03020201(02)	780051W352232N*	0.403	0.423	**+0.020**	0.442	0.464	**+0.022**	0.180	0.172	**-0.008**
03020201(02)	780051W352232N**	0.616	0.635	**+0.019**	0.750	0.793	**+0.043**	0.224	0.238	**+0.014**
03020201(02)	780051W352232N***	0.394	0.406	**+0.011**	0.426	0.439	**+0.013**	0.157	0.156	**-0.001**
03020201(02)	780013W352043N	0.513	0.521	**+0.008**	0.582	0.604	**+0.022**	0.188	0.208	**+0.020**
03040203	790102W343559N *	0.295	0.282	**-0.013**	0.302	0.288	**-0.013**	0.072	0.074	**+0.002**
03040203	790102W343559N **	0.379	0.334	**-0.045**	0.409	0.360	**-0.049**	0.165	0.178	**+0.013**
03040203	790102W343559N ***	0.306	0.286	**-0.020**	0.315	0.294	**-0.021**	0.082	0.086	**+0.003**
Texas										
12010005	933659W304031N	0.494	0.490	**-0.004**	0.566	0.561	**-0.006**	0.206	0.205	**-0.001**
12010005	935525W304104N	0.489	0.489	+0.000	0.553	0.553	+0.000	0.191	0.191	+0.000
12010005	940714W300838N	0.575	0.566	**-0.009**	0.667	0.653	**-0.014**	0.193	0.191	**-0.002**
12020003	940714W300838N	0.449	0.452	**+0.003**	0.458	0.461	**+0.003**	0.038	0.039	**+0.001**
12020006	940714W300838N	0.567	0.578	**+0.011**	0.865	0.922	**+0.057**	0.378	0.392	**+0.015**
12020007	940714W300838N	0.534	0.556	**+0.022**	0.657	0.709	**+0.053**	0.260	0.280	**+0.021**
12040102	953124W295936N	0.428	0.404	**-0.024**	0.746	0.791	**+0.045**	0.499	0.548	**+0.049**

* HWM (Sep28-Oct9,2016) **HWM (Oct5-Oct20, 2016) ***Oct 9,2016 at 03:00 (snapshot)

6. Conclusion

This study investigated how stage-varying Manning's roughness affects OWP HAND-FIM synthetic rating curves (SRCs) and FIM predictions. By correcting the channel-bed datum and calibrating roughness across five recurrence-based stage zones, median discharge error at held-out rating-curve points fell from 74.5% to 3.6%, with 98% of gauges showing improvement. However, the relationship between roughness and stage varied across gauges. Predictability of the roughness coefficient at ungauged locations was low, with R^2 ranging from 0.19 to 0.31, with the exception of the baseflow parameter, which showed high predictability ($R^2 > 0.8$). While, evaluating FIM using the calibrated roughness coefficient through a simplified n branch-transfer method produced mixed results relative to baseline FIM models across multiple benchmarks, suggesting potential for this approach. Across all three analyses, a wider study area is needed to better characterize the drivers of stage-roughness variability, the predictors of the roughness coefficient, and the reliability of FIM benefits at scale. Uncertainty compounds across the workflow: rating curves rarely capture the largest discharges, the DEM lacks bathymetry, only 128 gauges were calibrated, and flood mapping relies on NWM v3 discharge without data assimilation. Benchmarks also cover only high-magnitude events. This is a proof of concept, not a finalized method. Next steps are to test the propagation rule directly, complete the Hurricane Florence benchmarks, and extend calibration toward a CONUS-wide correction. Restricting FIM evaluation to calibrated reaches, rather than the entire benchmark, will yield more direct estimates of performance. The broader lesson is that a correction validated at the point scale does not automatically carry to the map scale, and the quality of its spatial extension can matter as much as the correction itself.

Acknowledgements: We thank Reza Khanbilvardi, Professor at The City College of New York; Humberto Vergara, Professor at the University of Iowa; Alvar Escriva-Bou of the University of California, Davis; Assefa Melesse, Professor at Florida International University; and Fabio Castelli, Professor at the University of Florence, for their academic guidance. We also thank Supath Dhital and Dipsikha Devi of The University of Alabama for their guidance on the use of FIMbox, FIMeval, and FIMbench.

We are grateful to the Water Prediction Innovators Summer Institute Program for organizing the program and supporting this work. Special thanks go to Nana Oye Djan and Megan Vardaman for their excellent coordination throughout the Summer Institute.

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA.

Claude Sonnet 5 was used to assist with grammatical editing and coding. All factual content, results, and interpretations were reviewed and verified by the authors.

Supplementary Materials: All the processes and codes are available in the following GitHub repository: https://github.com/sdmlua/fimbox

References

[1] A. D. Nobre *et al.*, "Height Above the Nearest Drainage – a hydrologically relevant new terrain model," *J. Hydrol.*, vol. 404, no. 1–2, pp. 13–29, Jun. 2011, doi: 10.1016/j.jhydrol.2011.03.051.

[2] NOAA-OWP, *NOAA-OWP/inundation-mapping Wiki*. GitHub. Accessed: Jun. 19, 2026. [Online]. Available: https://github.com/NOAA-OWP/inundation-mapping/wiki

[3] X. Zheng, D. G. Tarboton, D. R. Maidment, Y. Y. Liu, and P. Passalacqua, "River Channel Geometry and Rating Curve Estimation Using Height above the Nearest Drainage," *JAWRA J. Am. Water Resour. Assoc.*, vol. 54, no. 4, pp. 785–806, Aug. 2018, doi: 10.1111/1752-1688.12661.

[4] L. K. Read, D. N. Yates, J. M. McCreight, A. Rafieeinasab, K. Sampson, and D. J. Gochis, "Development and evaluation of the channel routing model and parameters within the National Water Model," *JAWRA J. Am. Water Resour. Assoc.*, vol. 59, no. 5, pp. 1051–1066, Oct. 2023, doi: 10.1111/1752-1688.13134.

[5] A. Baruah *et al.*, "Predicting synthetic rating curve adjustment factors with explainable machine learning for enhancing the United States operational flood inundation mapping framework," *J. Hydrol.*, vol. 662, p. 134086, Dec. 2025, doi: 10.1016/j.jhydrol.2025.134086.

[6] V. T. Chow, *Open-channel hydraulics*. New York: McGraw-Hill, 1959.

[7] G. Barinas, S. P. Good, and D. Tullos, "Continental Scale Assessment of Variation in Floodplain Roughness With Vegetation and Flow Characteristics," *Geophys. Res. Lett.*, vol. 51, no. 1, p. e2023GL105588, Jan. 2024, doi: 10.1029/2023GL105588.

[8] I. Garousi-Nejad, D. G. Tarboton, M. Aboutalebi, and A. F. Torres-Rua, "Terrain Analysis Enhancements to the Height Above Nearest Drainage Flood Inundation Mapping Method," *Water Resour. Res.*, vol. 55, no. 10, pp. 7983–8009, Oct. 2019, doi: 10.1029/2019WR024837.

[9] A. Baruah, R. Zarrabi, S. Cohen, J. M. Johnson, and R. McDermott, "Interpretable machine learning for predicting rating curve parameters using channel geometry and hydrological attributes across the United States," *Sci. Rep.*, vol. 15, no. 1, p. 44164, Dec. 2025, doi: 10.1038/s41598-025-27881-2.

[10] Y. Chen *et al.*, "Merging Remote Sensing Derived River Slope Datasets with High-Resolution Hydrofabrics for the United States," *Sci. Data*, vol. 12, no. 1, p. 1657, Oct. 2025, doi: 10.1038/s41597-025-05941-6.

[11] J. M. Johnson, D. Munasinghe, D. Eyelade, and S. Cohen, "An integrated evaluation of the National Water Model (NWM)–Height Above Nearest Drainage (HAND) flood mapping methodology," *Nat. Hazards Earth Syst. Sci.*, vol. 19, no. 11, pp. 2405–2420, Nov. 2019, doi: 10.5194/nhess-19-2405-2019.

[12] C. D. Rennó *et al.*, "HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia," *Remote Sens. Environ.*, vol. 112, no. 9, pp. 3469–3481, Sep. 2008, doi: 10.1016/j.rse.2008.03.018.
[13] M. A. Al Mehedi *et al.*, "Spatiotemporal Variability of Channel Roughness and its Substantial Impacts on Flood Modeling Errors," *Earths Future*, vol. 12, no. 7, p. e2023EF004257, Jul. 2024, doi: 10.1029/2023EF004257.
[14] "Surface Dynamics Modeling Lab," GitHub. Accessed: Jun. 15, 2026. [Online]. Available: https://github.com/sdmlua
[15] E. J. Kennedy, "Discharge ratings at gaging stations," U.S. Geological Survey, Techniques of Water-Resources Investigations Book 3, Chapter A10, 1984. doi: 10.3133/twri03A10.
[16] S. E. Rantz, "Measurement and computation of streamflow," U.S. Geological Survey, USGS Water-Supply Paper 2175, Vol. 2, 1982. doi: 10.3133/wsp2175.
[17] M. Wieczorek, S. Jackson, and G. (2018) Schwarz, "Select attributes for NHDPlus version 2.1 reach catchments and modified network routed upstream watersheds for the conterminous United States," *US Geol. Surv.*, vol. 10, p. F7765D7V, 2018.
[18] L. McKay, T. Bondelid, T. Dewald, J. Johnston, R. Moore, and A. Rea, "NHDPlus version 2: User guide," *US Environ. Prot. Agency*, vol. 745, 2012.
[19] R. A. Hill, M. H. Weber, S. G. Leibowitz, A. R. Olsen, and D. J. Thornbrugh, "The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States," *JAWRA J. Am. Water Resour. Assoc.*, vol. 52, no. 1, pp. 120–128, Feb. 2016, doi: 10.1111/1752-1688.12372.
[20] G. W. Abeshu, H.-Y. Li, Z. Zhu, Z. Tan, and L. R. Leung, "Median bed-material sediment particle size across rivers in the contiguous US," *Earth Syst. Sci. Data*, vol. 14, no. 2, pp. 929–942, 2022, doi: 10.5194/essd-14-929-2022.
[21] G. E. Schwarz and R. Alexander, "State Soil Geographic (STATSGO) Data Base for the Conterminous United States," Report 95–449, 1995. doi: 10.3133/ofr95449.

Chapter 6: Advancing ML and AI Frameworks for Enhanced Hydrologic Prediction

Basit Akinade[1], Ahmed Omar[2], Sanjeev Panta[3], Saddy Pineda-Castellanos[4], Mohammad Ali Javidian[5*], Sushant Mehan[6*]

[1]The University of Alabama; *baakinade@crimson.ua.edu*
[2]Texas A&M University-Corpus Christi; *aomar1@islander.tamucc.edu*
[3]University of Louisiana at Lafayette; *sanjeev.panta1@louisiana.edu*
[4]Utah State University; *saddy.pineda@usu.edu*
[4]Appalachian State University; *javidianma@appstate.edu*
[6]South Dakota State University; *sushant.mehan@sdstate.edu*
[*]Theme Leader

Abstract: The National Water Model (NWM), NOAA's operational physics-based hydrologic model, provides continental streamflow estimates but carries a systematic local bias, a consistent over- or under-estimation that varies by site and season and limits local use. We tested whether different machine learning algorithms (a simple recurrent network, a gated recurrent unit, a long short-term memory network, and a Transformer) can reduce this bias and asked where correction is most useful across three sites in three states (NC, VA, and SD). We trained these four sequence models under two setups: a residual setup that learns the difference between NWM discharge and observed United States Geological Survey (USGS) discharge and adds the learned correction back to NWM, and a direct setup that predicts observed discharge directly rather than a correction to NWM, with NWM still among its inputs. Records were split 70% for training, 15% for validation, and 15% for testing. We combined the models by simple averaging, error-weighted averaging, and constrained stacking, and tuned and scored each with time-ordered (walk-forward) cross-validation on a withheld recent block. The identical workflow was applied to three unregulated USGS gauges (no major upstream dams or diversions) spanning NWM skill, measured by the Kling-Gupta Efficiency (KGE), from 2010 to 2020 at hourly resolution: Watauga River (high skill), New River (moderate), and Little Spearfish Creek, a groundwater-fed karst spring (poor). Correction gains increased as NWM skill decreased: KGE improved by +0.12 at Watauga, +0.31 at New River, and +6.90 at Little Spearfish, with the highest corrected KGE reaching 0.83, 0.78, and 0.50 respectively. The residual setup performed best where NWM was reliable. Overall, ML correction generalized across regimes and added the most value where NWM skill was lowest, but at the karst spring it reached a ceiling set by driving information absent from the inputs.

1. Motivation

The National Water Model (NWM) is an essential tool for operational hydrology across the United States, producing streamflow estimates on the continental river network that support flood awareness, water supply, and research-to-operations workflows [1]. Because it is a physics-based model calibrated at regional scale rather than to each local gauge, its simulated streamflow can miss the timing of flood peaks, underestimate peak magnitude, or show systematic volume bias, and these biases differ from place to place and from season to season [1], [4]. Local decisions, such as small-basin flood response and reservoir operations, depend on the accuracy of simulated streamflow at the gauge, which is where a continental model is least constrained by local observations.

A practical response is to correct the model output after simulation, treating the physics-based estimate as a skillful first guess and using machine learning, algorithms that learn statistical patterns from data, to estimate and remove its local error. Error prediction and statistical post-processing have repeatedly improved streamflow forecasts without rebuilding the underlying physics [2], [3]. This motivates our central question: can sequence-based and ensemble machine learning (combining several models so their individual errors partly cancel) learn the local error of NWM against USGS observations and reduce it, and if so, at which sites does correction add the most value? Our three study sites deliberately include a groundwater-fed karst spring, where flow is driven by subsurface storage that NWM represents poorly, letting us probe the limits of the approach.

2. Objectives and Scope

This study pursued two objectives. 1) Build a reproducible machine-learning workflow that diagnoses the local bias in NWM streamflow, including its seasonal structure, and reduces it. 2) Identify where that correction adds the most value, measured as the gain in KGE (other metrics, including NSE, logNSE, and PBIAS, are reported in **Table 4** and **Figure 6**), across sites with contrasting NWM skill, using leak-free, time-ordered validation so the reported gains are reliable. By adding value, we mean a measurable improvement in agreement between corrected and observed streamflow.

We bounded the work to three unregulated USGS gauges (no major upstream dams or diversions) chosen to span NWM skill, measured by KGE, over 2010 to 2020 at hourly resolution. Model inputs were NWM streamflow (m^3/s), ERA5 (the ECMWF fifth-generation reanalysis [18]) total precipitation (mm) and volumetric soil moisture (m^3/m^3), and cyclic encodings of time; observed USGS streamflow (m^3/s) served only as the training target. We hypothesized that the residual setup would perform better where NWM is skillful and the direct setup where it is not, because a residual model only needs to learn a small correction when the physics is already close, whereas a direct model must reconstruct the whole hydrograph and is preferable only when NWM adds little information. We further expected that combining complementary base models (ensembling) would give more consistent accuracy across metrics than any single model.

3. Previous Studies

Evaluations of the National Water Model (NWM) across thousands of gauges reveal valuable yet inconsistent skill, with the poorest performance observed in the central and eastern United States. This spatial variability necessitates localized correction [1]. Post-processing existing physical models with statistical or machine learning methods is a well-established solution. A study by Naser Neisary et al. [4] aligns with our research, as they corrected NWM streamflow in the regulated western United States using a machine learning framework incorporating reservoir storage and snow water equivalent. Related research focuses on predicting model errors to enhance runoff [2], comparing residual and direct Long Short-Term Memory (LSTM) post-processing techniques [3], and integrating multiple base learners through ensemble stacking [5]. Other investigations involve blending hydrological models with machine learning in dam-regulated basins [6, 7] and assimilating USGS observations within the NextGen framework [8]. A forty-year review confirms these advancements but highlights that hydrological post-processing is applied inconsistently and often presupposes a stable climate [9].

These studies present two notable gaps. Firstly, corrections are almost exclusively demonstrated in single, often regulated, basins [3], [4], [6], [7], leaving the relationship between correction efficacy and the underlying model's skill underexplored. Secondly, reported improvements typically rely on a single time split, and ensemble stacking frequently lacks protection against data leakage, the reuse of the

same data to fit and to score a model, which inflates apparent skill [17]. This matters because a blend that looks strong on one split may not generalize, so we guard against it with expanding-window walk-forward validation, a sealed test block, and stacking weights fit only on out-of-fold predictions. Our contribution is a controlled comparison across sites exhibiting a wide range of NWM skill, from high to low. By maintaining a fixed workflow and varying only the study site, we can quantify how the value of correction is influenced by baseline model skill. Our methodology draws upon deep learning (multi-layer neural networks) for rainfall-runoff modeling [10], constrained forecast combination and stacked generalization [11], [12], [13], streamflow-imputation reviews by Gao et al., Gill et al., and Hamzah et al. [14], [15], [16], and the critical requirement for time series validation in temporal order [17].

4. Methodology

4.1 Study area and data

We selected three unregulated USGS gauges, representative of the range of National Water Model (NWM) skill as quantified by the Kling-Gupta Efficiency (KGE) metric. The Watauga River near Sugar Grove, North Carolina (03479000) represents a flashy mountain stream where NWM skill is high. The New River near Galax, Virginia (03164000) exemplifies a large, slow-responding river with moderate NWM skill. Conversely, Little Spearfish Creek near Lead, South Dakota (06430850), a small, steady karst spring primarily fed by groundwater, exhibits poor NWM skill. A summary of the basic site characteristics is provided in **Table 1**.

For each gauge, hourly observed streamflow data from 2010 to 2020 were paired with NWM-simulated streamflow from the co-located river reach. The co-located reach was identified by its NWM reach identifier (COMID), which overlaps the USGS station. Paired data included ERA5 reanalysis total precipitation and volumetric soil moisture [18], along with cyclic encodings for hour, day of year, and month. It is important to note that the observed USGS streamflow served exclusively as the training target for the model and was not utilized as a model input.

Table 1. *Study-site characteristics. CV (coefficient of variation, standard deviation over mean) is lower for steadier flow.*

Attribute	Watauga 03479000	New River 03164000	Little Spearfish 06430850
Setting	Mountain stream (NC)	Large upland river (VA)	Karst spring creek (SD)
NWM reach (COMID)	19743430	6887572	5481901
Record	2010 to 2020	2010 to 2020	2010 to 2020
USGS median flow (cms)	3.88	50.8	0.61
USGS variability (CV)	1.81	1.05	0.29
USGS peak flow (cms)	349	1388	6.1
Imputed hours	3,687 (3.8%)	6,643 (6.9%)	3,251 (3.4%)

4.2 Dataset cleaning and imputation

Each record was aligned to a continuous hourly index and then converted to cubic meters per second. Short gaps in the observed series were filled using a benchmark approach. In this approach, known values were systematically removed in blocks and subsequently reconstructed. These reconstructed values were then compared against the withheld true values to assess the efficacy of various imputation

methods. We compared linear interpolation with three predictor-based methods (k-nearest neighbors, iterative multivariate imputation, and a random forest); linear interpolation was the most accurate at every gap length and was adopted at all sites (Section 5), because hourly streamflow is strongly autocorrelated over short gaps. Model inputs were standardized (z-scored) using statistics fit on training data only, and the strongly skewed streamflow and precipitation inputs were log-transformed as log(1+x) before scaling. Imputed hours were flagged and excluded from all loss and metric calculations, ensuring that models were trained and evaluated solely on genuine observations. The extent of imputation was limited, ranging between 3.4% and 6.9% of total hours (**Table 1**), which aligns with findings from streamflow-imputation reviews by Gao et al., Gill et al., and Hamzah et al. [14], [15], [16]. Here, gap length refers to the number of consecutive missing hours; the accuracy of each imputation method as a function of gap length is presented in the Results section.

4.3 Bias diagnosis and input selection

We characterized the raw NWM error in simulated streamflow against USGS observations using the Kling-Gupta Efficiency and its three components: correlation r, variability ratio, and volume ratio, together with percent bias [19]. This approach distinguishes timing errors from variance and volume errors, thereby identifying specific parts of the hydrograph that the NWM reproduces poorly. Candidate predictors were ranked by correlation, mutual information, and permutation importance. All three rankings agreed: NWM streamflow, recent precipitation, and soil moisture ranked highest; consequently, only these dynamic inputs (along with time encodings) were retained. This small feature set maintains model simplicity and mitigates the risk of overfitting.

4.4 Seasonality

To understand when the NWM error is largest, we computed a monthly climatology of the bias, summarized by meteorological season. The bias proved strongly seasonal (Section 5): NWM over-predicts autumn flow at Watauga and, most severely, summer flow at the karst spring, where the timing correlation is near zero in every season. The lesson is that a substantial part of the error is predictable from the calendar, which is why we supplied cyclic encodings of hour, day of year, and month so the models can condition their corrections on season.

4.5 Modeling framework

For each gauge we trained four sequence models: a simple recurrent network (RNN, a type of recurrent neural network), a gated recurrent unit (GRU) [20], a long short-term memory network (LSTM) [21], [10], and a Transformer encoder [22]. These four span the standard progression of sequence architectures, from a simple recurrent baseline through gated networks that handle long-range dependencies (GRU and LSTM) to an attention-based Transformer, so we could test whether added architectural complexity helps for this task. Each model reads a lookback window of L past hours of inputs and predicts a single value at the current hour. We used two setups. In the residual setup, the model predicts the NWM error (observed minus simulated streamflow) and the correction is obtained by adding this prediction back to NWM (Equation 1). In the direct setup, the model predicts the streamflow itself rather than a correction to NWM (Equation 2); NWM streamflow remains one of the inputs in both setups, so 'direct' refers to the target, not to dropping NWM.

$$\hat{q}_{corr}(t) = q_{NWM}(t) + f_\theta(X_{t-L+1:t}), \qquad r(t) = q_{USGS}(t) - q_{NWM}(t) \tag{1}$$

$$\hat{q}_{corr}(t) = g_\phi(X_{t-L+1:t}), \qquad target\colon q_{USGS}(t) \tag{2}$$

In Equations 1 and 2, q̂(t) is corrected streamflow at hour t (cms), $q_{NWM}(t)$ and $q_{USGS}(t)$ are NWM and observed streamflow (cms), X is the matrix of scaled inputs over the lookback window, r(t) is the NWM error (cms), and f and g are the learned networks with parameters θ and φ. Both setups are trained by minimizing the mean squared error on the standardized target over observed hours only (Equation 3), where $\hat{y}_i$ and y_i are the predicted and observed standardized target for sample i (dimensionless) and N is the number of observed samples.

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_i (\hat{y}_i - y_i)^2 \tag{3}$$

The lookback length and network hyperparameters (hidden size, number of layers, dropout, learning rate, attention heads) were selected per model and setup with Optuna [23], an automated hyperparameter-search library that proposes and evaluates candidate settings to maximize a chosen score. Training used the Adam optimizer [24] with gradient clipping and early stopping, and the epoch that gave the highest validation KGE was retained.

4.6 Evaluation, cross-validation, ensembles, and uncertainty

Models were scored with the Kling-Gupta Efficiency (Equation 4) [19], the Nash-Sutcliffe Efficiency (Equation 5) [25], the Nash-Sutcliffe Efficiency of log-transformed flows to emphasize low flows (Equation 6), and percent bias (Equation 7).

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \qquad \alpha = \frac{\sigma_s}{\sigma_o},\ \ \beta = \frac{\mu_s}{\mu_o} \tag{4}$$

$$\mathrm{NSE} = 1 - \frac{\sum_i (s_i - o_i)^2}{\sum_i (o_i - \bar{o})^2} \tag{5}$$

$$\mathrm{logNSE} = 1 - \frac{\sum_i (\ln(s_i + \varepsilon) - \ln(o_i + \varepsilon))^2}{\sum_i (\ln(o_i + \varepsilon) - \overline{\ln(o + \varepsilon)})^2} \tag{6}$$

$$\mathrm{PBIAS} = 100 \times \frac{\sum_i (s_i - o_i)}{\sum_i o_i} \tag{7}$$

In Equations 4 to 7, s_i and o_i are simulated and observed streamflow for sample i (cms), ō is the mean observed streamflow (cms), *r* is the correlation between s and o, α is the ratio of their standard deviations, β is the ratio of their means (both dimensionless), and ε is a small positive constant (0.001 cms) that avoids the logarithm of zero. All three efficiencies equal 1 for a perfect match; percent bias is 0 when total simulated and observed volumes are equal. A single recommended model per site was chosen by the highest test-set KGE, and a robustness check based on the Moriasi criteria (log Nash-Sutcliffe Efficiency of at least 0.5 and absolute percent bias of at most 15 percent) flagged whether that model is reliable across both normal and low flows [26], [27].

To keep the scores reliable for time series, we replaced the single time split with expanding-window walk-forward cross-validation, nested inside a final untouched test block. We refer to the earlier portion of the record used for tuning as the development region, and to the final, most recent block, held aside and used only once for the reported scores, as the sealed test block. The split reserved the final 15 percent of each record as the sealed test block, with roughly 70 percent of the record for

training and 15 percent for validation. Within the development region the data were divided into five time-ordered folds. Each fold trained on all earlier data and validated on the next unseen block. The input scalers were refit within each fold, and a gap of L hours was left at every fold boundary, so no future information could influence an earlier prediction. Hyperparameters were selected to maximize the mean validation KGE across folds (Equation 8), and the spread across folds provided an error bar on each metric. The final model was then retrained on the whole development region and evaluated once on the sealed test block, which no model saw during training or tuning.

Within each setup, the base models were combined by three methods: equal-weight averaging, error-weighted averaging, and constrained stacking. Stacking learns a weighted average of the base predictions (Equation 9) in which the weights are non-negative and sum to one (Equation 10). This constrained form, following Granger and Ramanathan [11] and stacked generalization [13], keeps the blended prediction physically sensible (it cannot exceed the range of its inputs). The weights were fit by least squares (Equation 11) on out-of-fold predictions, that is, on each base model's outputs for validation blocks it never trained on, and then applied to the sealed test block, so the ensemble cannot inflate its own score by reusing training data. In our experiments, fitting the weights to maximize KGE tended to place almost all weight on a single model when the base models were correlated, so we used the least-squares fit instead.

$$w_k \geq 0, \qquad \sum_k w_k = 1 \tag{10}$$

$$w = \text{argmin} \sum \left(\sum_k w_k \ \hat{s}_k^{\text{OOF}} - o \right)^2 \tag{11}$$

In Equations 9 to 11, $\hat{s}_{stack}$ is the blended prediction (cms), $\hat{s}_k$ is the prediction of base model k (cms), w_k is its weight (dimensionless), $\hat{s}_k^{OOF}$ is its out-of-fold prediction, and o is the observed streamflow (cms). A flow-conditioned, split-conformal 90 percent band captures predictive uncertainty: on a held-out half of the test period, corrected-flow residuals are binned by predicted flow, and each bin's 5th and 95th percentiles set bounds that widen where errors are larger. Coverage is the fraction of observations falling inside the band.

5. Results

5.1 NWM skill before correction and event response

The three gauges encompass the full spectrum of NWM skill, as detailed in **Table 2**. Over the observed record, NWM skill, as measured by the Kling-Gupta Efficiency (KGE), is high at Watauga (0.76), moderate at New River (0.54), and notably poor at Little Spearfish (-1.06), where it performs worse than a mean-flow benchmark and exhibits a timing correlation of only 0.05.

Figure 1 presents the observed and simulated hydrographs, along with precipitation data, for a representative storm event at each site. This event is centered around the largest observed peak. At Watauga and New River, rainfall initiates a sharp increase in flow, which the NWM reproduces but with temporal shifts and underestimated magnitudes, indicating a correctable error. Conversely, at the karst spring, the same rainfall elicits minimal response. This is attributed to the flow being predominantly governed by slow groundwater storage, a process not represented by the NWM. Consequently, there is no discernible rainfall-driven signal to correct at this location.

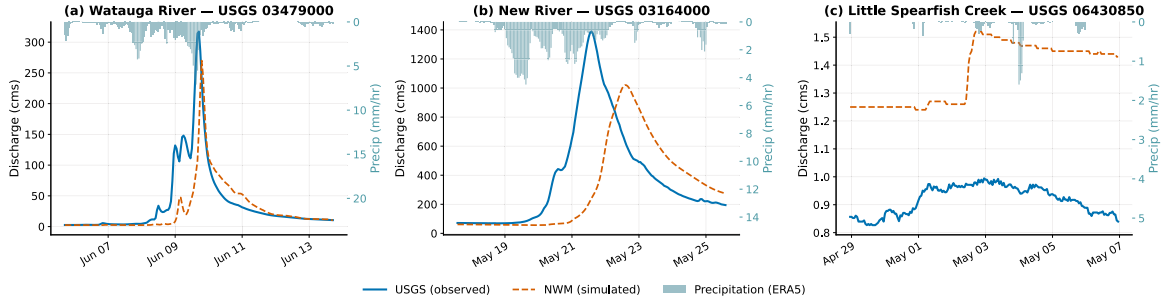


Figure 1. *Hourly hydrographs at a representative storm event for Watauga (a), New River (b), and Little Spearfish (c). Blue solid line: observed USGS streamflow; orange dashed line: NWM; light teal bars (inverted top axis): precipitation.*

Table 2. *Raw NWM skill before correction over the observed record. KGE, r, and NSE are higher-is-better (1 is perfect); PBIAS is 0 for no volume bias; alpha and beta are the variability and volume ratios.*

Metric	Watauga	New River	Little Spearfish
KGE	0.76	0.54	-1.06
r (timing)	0.79	0.61	0.05
NSE	0.62	0.33	-8.31
PBIAS	+4.5%	-10.6%	+23.6%
alpha (variability)	0.90	0.79	2.82
beta (volume)	1.05	0.89	1.24

For gap filling, linear interpolation proved to be the most accurate method across all gap lengths, successfully estimating 79% to 99% of masked test gaps against the held-out truth. This high accuracy is attributed to the strong autocorrelation of hourly streamflow, where proximal observed values provide more informative estimates than predictor-based approaches. Consequently, linear interpolation was adopted for all sites.

5.2 Seasonal bias, quantified and corrected

The National Water Model (NWM) exhibits a strong seasonal bias (**Table 3**). Specifically, the NWM over-predicts fall flow at Watauga and, more significantly, summer flow at Little Spearfish. At Little Spearfish, the timing correlation is near zero or negative across all seasons, indicating that the NWM's underlying physics does not accurately represent the karst system. **Figure 2** illustrates the per-season skill before and after correction. The correction improves the Kling-Gupta Efficiency (KGE) across all seasons at Watauga and New River. However, at Little Spearfish, the seasonal skill remains low, which is consistent with the absence of a groundwater signal in the model.

Table 3. *Seasonal NWM bias before correction: percent bias (PBIAS) and timing correlation (r) by season. Wat, NR, and LS denote Watauga, New River, and Little Spearfish.*

Season	Wat PBIAS	Wat r	NR PBIAS	NR r	LS PBIAS	LS r
DJF (winter)	+7%	0.75	-6%	0.66	+8%	-0.30
MAM (spring)	-6%	0.82	-22%	0.50	+8%	0.29
JJA (summer)	+2%	0.84	-8%	0.71	+53%	-0.11
SON (fall)	+24%	0.74	+2%	0.59	+21%	-0.27

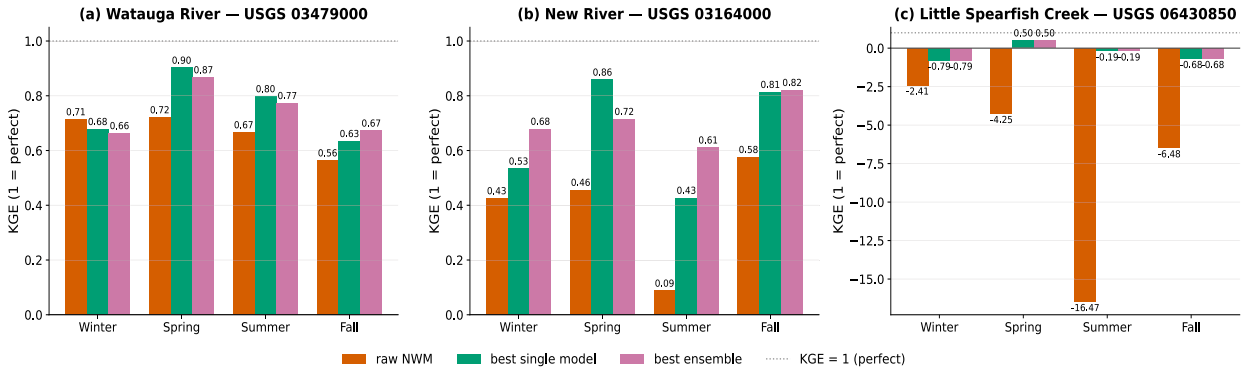


Figure 2. *Per-season KGE before and after correction for Watauga (a), New River (b), and Little Spearfish (c). Orange: raw NWM; green: best single model; pink: best ensemble.*

5.3 Model performance

The correction improved performance across all sites, with gains increasing as the NWM skill decreased (**Table 4**, **Figure 3**). The corrected KGE reached 0.83 at Watauga, representing a gain of +0.12 over the raw NWM value of 0.71 on the test block. Similarly, the KGE at New River improved to 0.78 (+0.31), and at Little Spearfish, it reached 0.50 (+6.90). The gain observed at New River is more than double that at Watauga. While initially separated by 0.24 KGE, the two responsive sites ultimately converged to within approximately 0.05 KGE, signifying a nearly fourfold reduction in the discrepancy. Watauga and New River meet the Moriasi robustness criteria; however, Little Spearfish does not (log Nash-Sutcliffe Efficiency of -3.4 and percent bias of +34 percent). This indicates that Little Spearfish is a less reliable fallback, and its substantial nominal gain primarily reflects the poor quality of its baseline performance. **Figure 3** presents a comparison of KGE and NSE for each model against the raw-NWM baseline. Most models outperform the baseline at Watauga and New River, whereas at Little Spearfish, only the direct-setup models achieve a positive KGE.

Table 4. *Recommended model at each site, on the sealed test block, with the cross-validated KGE (mean and spread across folds).*

Metric	Watauga (Transformer, residual)	New River (LSTM, residual)	Little Spearfish (LSTM, direct)
Raw NWM KGE (test)	0.71	0.47	-6.40
Corrected KGE	0.83	0.78	0.50
Cross-validated KGE (mean, std)	0.78 ± 0.02	0.70 ± 0.10	0.05 ± 0.21
NSE	0.69	0.64	-4.48
logNSE	0.80	0.77	-3.44
PBIAS	+1.2%	-5.9%	+33.9%
Passes Moriasi bands	yes	yes	no

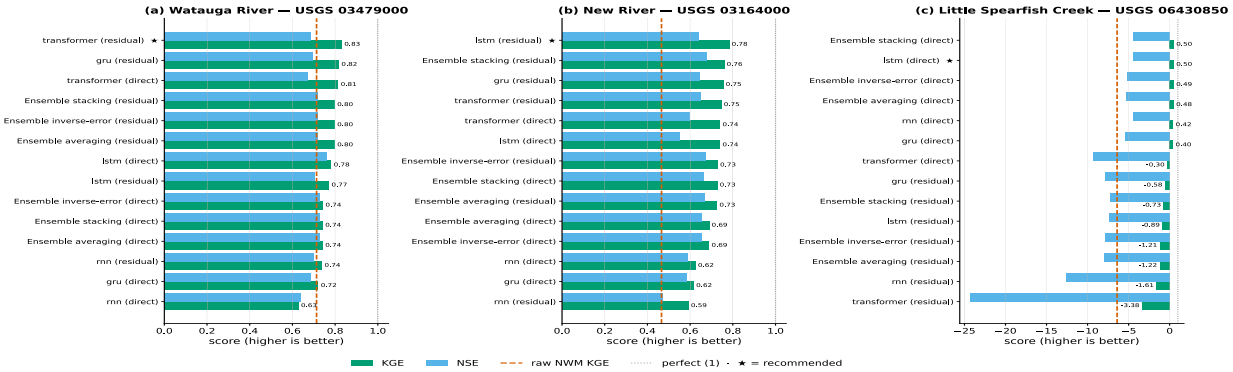


Figure 3. *Model comparison by site: KGE (green) and NSE (blue) for every model, with the raw-NWM KGE marked by the dashed line and the recommended model starred. Watauga (a), New River (b), Little Spearfish (c).*

5.4 Residual versus direct setup

The two experimental setups performed as hypothesized (**Figure 4**). At the Watauga and New River sites, where the National Water Model (NWM) is reliable, correcting its error (the residual setup) resulted in performance that matched or exceeded direct streamflow prediction. Conversely, at Little Spearfish, where the NWM exhibits a poor signal, the residual setup yielded strongly negative Kling-Gupta Efficiency (KGE) values. This outcome suggests an absence of discernible errors for correction, with only the direct prediction setup achieving positive skill. These findings support a straightforward principle: utilize NWM error correction in contexts where the model demonstrates reliability and employ direct streamflow prediction where its reliability is compromised.

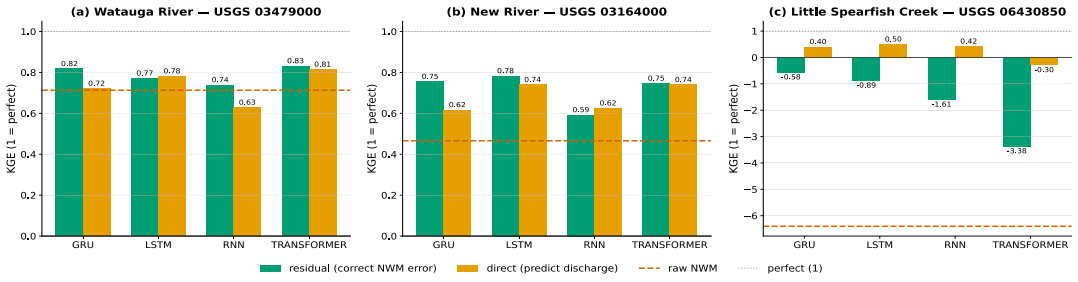


Figure 4. *Residual (green) versus direct (orange) setup by model, as KGE, with the raw-NWM KGE marked by the dashed line. Watauga (a), New River (b), Little Spearfish (c).*

5.5 Largest-event hydrographs

During the largest observed event (**Figure 5**), the corrected model more accurately tracked the observed peak at Watauga and New River than the raw National Water Model (NWM), demonstrating improvements in both timing and magnitude, which remained within the 90 percent uncertainty band. At Little Spearfish, the event was minor and exhibited a weak correlation with rainfall, thus the correction provided minimal additional benefit.

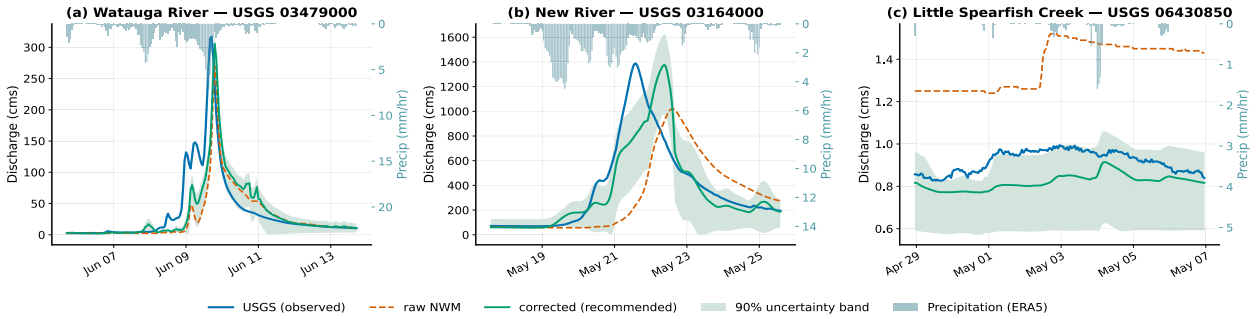


Figure 5. *Corrected hydrographs at the largest observed event for Watauga (a), New River (b), and Little Spearfish (c). Blue: observed USGS; orange dashed: raw NWM; green: corrected model; green shading: 90 percent band; teal bars (inverted): precipitation.*

5.6 Full metric comparison

The comprehensive performance of all models across various metrics is presented in **Figure 6**. The Watauga and New River sites consistently exhibit favorable model ranks across most metrics, while the Little Spearfish site is characterized by predominantly unfavorable ranks, indicating the unreliability of even its best-performing model. It is notable that single models excelling in one headline metric rarely demonstrate overall superior performance. For instance, at both responsive sites, the top-KGE model performs poorly on error measures such as MAE and low-flow RMSE. Conversely, ensemble models consistently rank near the top across all metrics simultaneously. This balanced performance across a full spectrum of metrics, rather than optimizing for one at the expense of others, underscores the practical advantage of using ensemble modeling approaches.

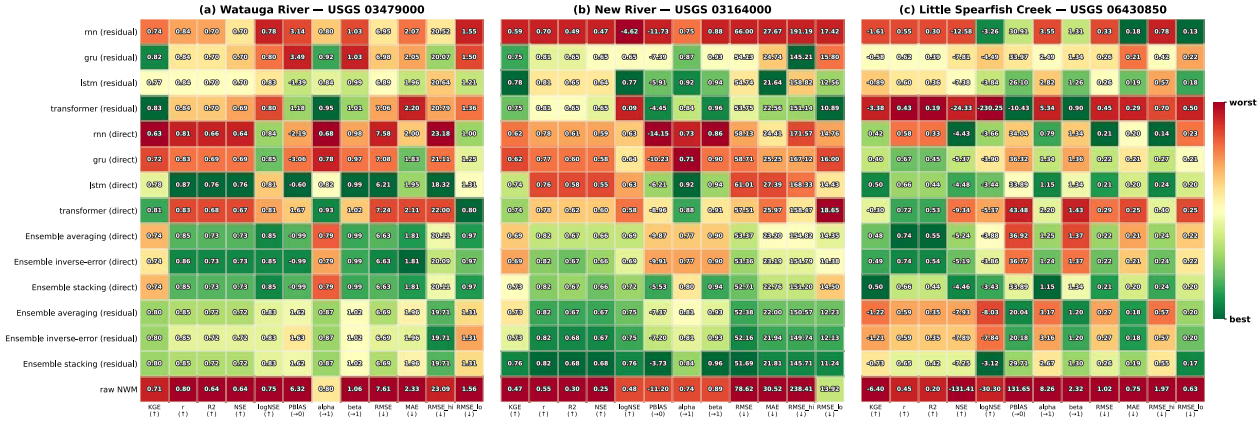


Figure 6. *Model scorecard across all metrics for the three sites. Each cell is colored by the model's rank on that metric (green best, red worst); the bottom row is raw NWM. Watauga (a), New River (b), Little Spearfish (c).*

6. Limitations and Future Work

The compact feature set facilitates transferability but omits crucial variables, such as deeper groundwater states, which are decisive at karst sites. The study encompasses three gauges, sufficient to establish a skill gradient but insufficient for generalization across the full spectrum of hydrologic settings served by the National Water Model (NWM). The NWM's operational domain now extends beyond the continental United States to include Alaska, Hawaii, and Puerto Rico. Future work will involve testing the pipeline at additional gauges, including sites exhibiting intermediate skill, and exploring inputs such as antecedent multi-month storage indices, which could enhance estimates for karst-influenced rivers. By design, the pipeline is station-adaptable; therefore, extending its application to a new gauge merely necessitates a clean dataset conforming to the shared schema.

7. Conclusion

Across three contrasting rivers, machine-learning correction of National Water Model (NWM) streamflow demonstrated significant value, particularly in instances where NWM skill, as measured by the Kling-Gupta Efficiency (KGE), was initially low. Corrected KGE values improved by +0.12, +0.31, and +6.90 from high, moderate, and low-skill baselines, respectively. This correction narrowed the skill gap between the two responsive sites from approximately 0.24 to 0.05 KGE.

Two practical guidelines emerged from this study. First, the correction method should be matched to the baseline NWM skill: NWM output should be corrected where it demonstrates reliability, whereas streamflow should be predicted directly when NWM reliability is low. Second, models that perform poorly should be excluded from ensembles by assigning them negligible weights. The hyperparameter search consistently indicated a preference for a short memory window (approximately 6 hours) for the flashy Watauga River and a longer window (approximately 72 hours) for the larger, slower New River, aligning with their respective hydrological response times to rainfall.

The credibility of these findings is supported by the walk-forward cross-validation, which showed widening fold-to-fold error bars as site complexity increased. Furthermore, the performance on the sealed test block matched the cross-validated score, indicating the absence of data leakage. The Little Spearfish River site represents a limitation of this approach. In this case, with minimal NWM signal to refine, the model was required to predict streamflow independently. Even a robust direct prediction is constrained by groundwater storage, which cannot be observed through surface inputs. This limitation stems from the available information, not the model itself, as no amount of training can recover a signal absent from the input data.

Acknowledgements: We thank our theme leaders, Dr. Mohammad Ali Javidian (Appalachian State University) and Dr. Sushant Mehan (South Dakota State University), and CUAHSI for program support. We also thank our academic advisors for their guidance: Dr. Amobichukwu C. Amanambu (The University of Alabama), Dr. Li Chen (University of Louisiana at Lafayette), Dr. Sarah E. Null (Utah State University), and Dr. Tianxing Chu (Texas A&M University-Corpus Christi).

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA.

Claude Opus 4.8 (Anthropic, July 2026) was used to assist with debugging portions of the analysis code and to correct grammar and clarity in the text, as the authors included non-native English speakers. All factual content, interpretations, citations, and code were reviewed, tested, and verified by the authors, who remain fully responsible for the work.

Supplementary Materials: The clean hourly datasets (aligned, unit-converted, gap-flagged streamflow, meteorology, and time encodings for the three gauges) and the full pipeline code (data preparation, models, training, walk-forward cross-validation, ensembles, and figure generation) are openly available. The datasets are archived on HydroShare [Link to HydroShare] and the code is in the project repository [Link to GitHub]. Imputed hours are flagged so that all training and scoring use genuine observations only.

References

[1] B. Cosgrove, D. Gochis, T. Flowers, A. Dugger, F. Ogden, et al., "NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling," JAWRA J. Amer. Water Resour. Assoc., vol. 60, no. 2, pp. 247-272, 2024, doi: 10.1111/1752-1688.13184.

[2] H. Han and R. R. Morrison, "Improved runoff forecasting performance through error predictions using a deep-learning approach," J. Hydrol., vol. 608, art. 127653, 2022, doi: 10.1016/j.jhydrol.2022.127653.

[3] S. Sharma, G. R. Ghimire, and R. Siddique, "Machine learning for postprocessing ensemble streamflow forecasts," J. Hydroinform., vol. 25, no. 1, pp. 126-139, 2023, doi: 10.2166/hydro.2022.114.

[4] S. Naser Neisary, R. C. Johnson, M. S. Alam, and S. Burian, "A post-processing machine learning framework for bias-correcting National Water Model outputs by accounting for dominant streamflow drivers," Environ. Model. Softw., vol. 190, art. 106459, 2025, doi: 10.1016/j.envsoft.2025.106459.

[5] H. Tyralis, G. Papacharalampous, and A. Langousis, "Super ensemble learning for daily streamflow forecasting," Neural Comput. Appl., vol. 33, pp. 3053-3068, 2021, doi: 10.1007/s00521-020-05172-3.

[6] H. Solanki, U. Vegad, A. Kushwaha, and V. Mishra, "Improving streamflow prediction using multiple hydrological models and machine learning methods," Water Resour. Res., vol. 61, art. E2024wr038192, 2025, doi: 10.1029/2024WR038192.

[7] J. Liu, X. Yuan, J. Zeng, Y. Jiao, Y. Li, L. Zhong, and L. Yao, "Ensemble streamflow forecasting over a cascade reservoir catchment with integrated hydrometeorological modeling and machine learning," Hydrol. Earth Syst. Sci., vol. 26, no. 2, pp. 265-278, 2022, doi: 10.5194/hess-26-265-2022.

[8] E. Foroumandi, H. Moradkhani, W. F. Krajewski, and F. L. Ogden, "Ensemble data assimilation for operational streamflow predictions in the Next Generation (nextgen) framework," Environ. Model. Softw., vol. 185, art. 106306, 2025, doi: 10.1016/j.envsoft.2024.106306.

[9] M. Troin, R. Arsenault, A. W. Wood, F. Brissette, and J.-L. Martel, "Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years," Water Resour. Res., vol. 57, art. E2020wr028392, 2021, doi: 10.1029/2020WR028392.

[10] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks," Hydrol. Earth Syst. Sci., vol. 22, no. 11, pp. 6005-6022, 2018, doi: 10.5194/hess-22-6005-2018.

[11] C. W. J. Granger and R. Ramanathan, "Improved methods of combining forecasts," J. Forecast., vol. 3, no. 2, pp. 197-204, 1984, doi: 10.1002/for.3980030207.
[12] D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241-259, 1992, doi: 10.1016/S0893-6080(05)80023-1.
[13] L. Breiman, "Stacked regressions," Mach. Learn., vol. 24, no. 1, pp. 49-64, 1996, doi: 10.1007/BF00117832.
[14] Y. Gao, C. Merz, G. Lischeid, and M. Schneider, "A review on missing hydrological data processing," Environ. Earth Sci., vol. 77, art. 47, 2018, doi: 10.1007/s12665-018-7228-6.
[15] M. K. Gill, T. Asefa, Y. Kaheil, and M. Mckee, "Effect of missing data on performance of learning algorithms for hydrologic predictions," Water Resour. Res., vol. 43, no. 7, art. W07416, 2007, doi: 10.1029/2006WR005298.
[16] F. B. Hamzah, F. Mohd Hamzah, S. F. Mohd Razali, O. Jaafar, and N. Abdul Jamil, "Imputation methods for recovering streamflow observation: A methodological review," Cogent Environ. Sci., vol. 6, no. 1, art. 1745133, 2020, doi: 10.1080/23311843.2020.1745133.
[17] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," Inf. Sci., vol. 191, pp. 192-213, 2012, doi: 10.1016/j.ins.2011.12.028.
[18] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, et al., "The ERA5 global reanalysis," Q. J. R. Meteorol. Soc., vol. 146, no. 730, pp. 1999-2049, 2020, doi: 10.1002/qj.3803.
[19] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and NSE performance criteria," J. Hydrol., vol. 377, no. 1-2, pp. 80-91, 2009, doi: 10.1016/j.jhydrol.2009.08.003.
[20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. EMNLP, 2014, pp. 1724-1734, doi: 10.3115/v1/D14-1179.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Adv. Neural Inf. Process. Syst. (neurips), vol. 30, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762
[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD, 2019, pp. 2623-2631, doi: 10.1145/3292500.3330701.
[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2015. [Online]. Available: https://arxiv.org/abs/1412.6980
[25] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I: A discussion of principles," J. Hydrol., vol. 10, no. 3, pp. 282-290, 1970, doi: 10.1016/0022-1694(70)90255-6.
[26] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," Trans. ASABE, vol. 50, no. 3, pp. 885-900, 2007, doi: 10.13031/2013.23153.
[27] D. N. Moriasi, M. W. Gitau, N. Pai, and P. Daggupati, "Hydrologic and water quality models: Performance measures and evaluation criteria," Trans. ASABE, vol. 58, no. 6, pp. 1763-1785, 2015, doi: 10.13031/trans.58.10715.

Chapter 7: Toward Better Flood Risk Communication: Applying a Diagnostic Framework for Evaluating and Redesigning FIM Visualizations

Dominic Kyei[1], Sohail Tufail[2], Kelsey McDonough[3*], Sagy Cohen[4*]

[1]The University of Kansas; *dkyei@ku.edu*
[2]University of Memphis; *satufail@memphic.edu*
[3]The Water Institute; *mcdonough.kelseyr@gmail.com*
[4]The University of Alabama; *sagy.cohen@ua.edu*
*Theme Leader

Abstract: Flood Inundation Maps (FIMs) are increasingly used to communicate flood extent, depth, timing, and hazard conditions, but their value depends on whether the displayed information can be interpreted clearly and translated into practical understanding. The study evaluates selected FIM platforms as communication interfaces rather than only technical map products. Using the Diagnosing, Assessing, Redesigning, Testing, and Synthesizing (DARTS) framework, survey data were collected from participants in the 2026 Summer Institute. The platforms assessed included public facing FIMs, including the NOAA National Water Prediction Service (NWPS) Emulator and EarlyFlows, and credentialed platforms, including FloodID and HurrEvac. Participants evaluated each platform using 1-5 rating grids, short-answer questions, and screenshots. Qualitative responses were analyzed using NVivo thematic analysis, while rating-grid results were synthesized to compare platform performance. Across four widely used flood inundation mapping platforms, users consistently struggled to interpret flood information despite navigating the interfaces with relative ease, indicating that the principal limitation of current FIMs lies in risk communication rather than usability. Common challenges included technical language, unclear visual design, poor communication of uncertainty, and confusing legends and map layers. Coded responses also identified recurring needs for clearer main messages, plain-language explanations, improved layer and color interpretation, clearer forecast timing and limitation guidance, and stronger impact communication. These findings were translated into prototype-informed NWPS design targets, including recommendations for simpler default layer organization, guided interpretation support, and an Impact Lens concept to indicate potentially inundated roadways. The study applies an evidence-based framework for redesigning flood inundation maps to achieve effective risk communication by linking user perceptions to specific interface design recommendations.

1. Motivation

Flood Inundation Maps (FIMs) are important tools for communicating where flooding may occur, how deep water may be, and when flood conditions may change [1]. In operational settings, FIMs help translate hydrologic model outputs into spatial information that can support situational awareness, warning interpretation, and response planning [1], [2]. However, a map that accurately shows flood extent is not automatically an effective risk communication product [3]. Users must also understand what the displayed information means, including whether the data displayed represents current or forecast flooding, what limitations apply, and any real-world impacts, such as flooded roads or buildings, that may be associated with the mapped flooding [4].

The communication challenge is especially important for web-based FIM interfaces. These platforms often combine hydrographs, forecast layers, gauges, legends, basemaps, warnings, model guidance, and supporting overlays in one singular interface [1], [3]. This technical richness can be valuable, but it can also create an interpretation challenge when colors, labels, layer names, timing information, or uncertainty cues are difficult to understand [1], [5]. Prior work on flood-risk information has emphasized that design choices affect how users interpret risk [6], compare alternatives forecast information [2], and make decisions under uncertainty [4], [2]. This study responds to that challenge by treating FIMs as communication interfaces.

2. Objectives and Scope

The objectives of this study were to develop and apply a Diagnosing, Assessing, Redesigning, Testing, and Synthesizing (DARTS) framework for evaluating FIM visualization clarity, interpretation burden, impact communication, and actionability; to identify recurring visualization strengths and weaknesses, as well as similarities and differences across selected public and credentialed platforms; and to translate diagnostic findings into practical NWPS Emulator design targets.

Specifically, the study aims to answer the following questions:

1. What visualization strengths and weaknesses appear in public facing and credentialed FIMs domains?
2. How do public facing and credentialed FIM platforms differ in communication and usability?
3. How can diagnostic findings be translated into practical visualization targets for the NWPS Emulator?

3. Previous Studies

For Flood Inundation Maps (FIMs) to effectively communicate risk, they must answer three questions: whether visuals improve understanding of risk information [7], [8], whether its design features shape the intended interpretation [7], [1], [9], and whether these visuals affect trust, emotions and illicit protective action [10]. This section explores how flood inundation maps can serve these purposes. Section 2.1 reviews the visualization principles within the broader risk communications literature, section 2.2 addresses how users experience current FIMs and section 2.3 considers what these findings imply for decision-support designs.

3.1 Risk Communication and Visual Interpretation

Risk communication research emphasizes that information is most useful when it is credible, understandable, relevant, and connected to user decisions [11], [12], [6]. Visualization is one of the most powerful ways for achieving that. Studies show that people process visual information faster than text, and a well-designed FIM can communicate complex flood information to users within seconds [13], [9], [5]. FIMs use several design features such as color, symbols, layouts, and legends to serve communication purposes. However, these features carry consequences for how users perceive, interpret, and ultimately act on the risk information presented to them.

Visual design choices in FIM platforms have direct effects on user comprehension. Colors and symbols convey information about inundated locations. Colors, in particular, function as a primary risk signal for most users. Studies have indicated that people interpret color according to deeply held cognitive association [5]. For example, people associate red and deep brown as a maximum danger,

whereas green signals safety. As such, interfaces which do not align with these associations can force users into additional interpretive effort that slows comprehension and increases the likelihood of misreading risk information [1]. Notably, many current FIM platforms, including the National Water Prediction Service (NWPS), denote green simultaneously to represent no-flood conditions [14]. This clusters visually similar elements across categorically different risk levels, creating a color environment in which the meaning of any individual green element cannot be inferred without consulting the legend.

Another design choice is regarding the use of symbols and legends. Different FIMs use triangles, squares, and circles, to represent gauged locations, forecast points and zones across the same interface. When these symbols are rendered in similar sizes and colors on a dense basemap, users cannot reliably distinguish between them without explicit legend guidance. Even so, many FIM platforms position legends in collapsed panels that require user interaction to access, embed them within multi-level layer menus that require navigation to find, or update them inconsistently as layers are toggled on and off [7], [4], [1]. These irregularities in FIM designs could be problematic and ineffective at the same time, requiring simplifying visual elements and using clear, distinguishable symbols and colors to enhance comprehension by reducing cognitive load and directing attention to essential information [2], [5].

Based on this, we hypothesize these,

H1a: FIM platforms that display a high density of overlapping colors will produce visual confusion regarding flood risk information.

H1b: FIM platforms that use symbols of similar size, shape, and color for categorically different risk features will lead users to misidentify the type and severity of the flood hazard being displayed.

H1c: FIM platforms where critical risk information is embedded within collapsed or inaccessible legends will be perceived as significantly more difficult to interpret than platforms where legend information is persistent and immediately visible.

3.2 FIMs and Users Perceptions

People's initial perception of a FIM can determine how well they consume the flood information it contains [2], [15]. For example, when a user opens a FIM for the first time, under stressed conditions and limited time, the initial interface can create an immediate impression that shapes everything that follows. FIM platforms with clustered interfaces, where layers of information compete simultaneously for attention, can overwhelm users before they can locate the most critical risk message [1]. This can be common among users who have no prior training or bring no specialist knowledge on the platform.

Credibility and trust is a highly discussed domain in flood risk communication When users experience cognitive overload, it may lead to trust and credibility challenges [11], [12]. Studies assert that users are more likely to engage and act on risk information from sources they perceive as credible and authoritative [10], [12]. Official government platforms like the NWPS could gain initial user trust however, this trust and credibility could be maintained when information about uncertainty is properly communicated to FIM users. In addition, when a user perceives FIM interface as confusing, users can interpret their difficulties as a signal that the flood risk situation is ambiguous rather than urgent [9], [16]. Moreso, when FIMs present a single deterministic flood extent as though they were a confirmed boundary, any deviation between the observed and forecasted scenarios can create a false impression

[5]. This requires the use of simplified language as well as different language designs to ensure that users can understand and receive the right messages when needed.

On this basis, we hypothesize the following;

H2a: FIMs designed with less visual clusters, easy navigation extent will heighten user first impression.

H2b: FIMs designed with clear flood extent and uncertainty communication can build trust and credibility leading to improved understanding.

3.3 Flood Inundation Mapping (FIMs) and Decision Support

Flood inundation maps (FIMs) are valuable tools for communicating flood risks to users. However, their effectiveness depends on users' ability to make informed protective decisions when interacting with them. Protective action theories reveal that users are more likely to take actions when flood risk information is paired with explicit actions guidance rather than presenting information about risk alone [17]. This infers that FIMs should be designed to be actionable. For example, FIMs should communicate inundated road networks as well as safe zones to aid user decisions about evacuation and travel. Surprisingly, many FIMs lack this actionability feature.

The distinction between different FIM platforms can also guide the understanding of how current FIMs can support protective action. Many credentialed FIMs are designed with features necessary to meet specific needs. Some are designed for experts such as meteorologists and hydrologists at forecasting centers. Examples include FloodID and HurrEvac which track storms and hurricane paths respectively. Due to these specialized client needs, such FIM platforms could be designed with huge technical jargons, and complex language. According to Gerst, et al. [13] many FIMs attempt to serve all possible users with a single interface, which often defaults to an expert-focused design that overwhelms non-experts.

We can hypothesize that;

H3a: FIMs designed with technical words, jargons and abbreviations without clear definitions will be perceived as a barrier to understanding flood risk.

H3b: Responders will express commendation towards FIMs which demonstrate actionability

3.4 Diagnostic Framework for FIM Evaluation

To address FIM visualization challenges and to develop prototype FIM products, many researchers have resorted to several frameworks. Some of these frameworks include the Internalization, Distribution, Explanation, and Action (IDEA) framework [18], The Integrated Crisis Mapping (ICM) model [9] and Standard User-Centered Design (UCD) framework [13]. While these frameworks could be useful in their own ways, their applicability to the current study could be limited. Each framework above relies on subjective and user centered assessments to propose a redesign. This could be problematic given that subjective assessments alone could expose the researchers to make biased decisions [19]. We employ the Diagnosing, Assessing, Redesigning, Testing, and Synthesizing (DARTS) framework to provide unique, research-backed evidence for specific redesign changes [3], [13]. In the diagnostic stage, existing visualization and risk communication literature is reviewed to identify recurring design challenges that limit FIM usability for users, and a working hypothesis is developed to guide the assessment that follows. In the Assessment stage, a structured set of

questionnaires is applied across selected FIM platforms to allow respondents to provide evidence about how they perceive, interpret, and interact with each platform's visual design. In the Redesign stage, findings from the assessment are synthesized with cross-disciplinary risk communication evidence to produce a set of evidence-based visualization targets to inform a redesign prototype. In the Testing and Synthesis stage, the redesigned prototype is evaluated with several users and experts whose feedback is then used to refine the design further before findings are synthesized into transferable design principles for the broader FIM community [1], [3]. This framework adopts a feedback loop at each stage to ensure subjectivity is reduced.

4. Methodology

4.1 Study Design

We built our study off on the work by Kandel, et al. [1]. We intended to understand how different respondents will express challenges and capabilities across existing FIM. For example, we adopted the same study questionnaire as used in their study. However, we included Likert-scale questions as an improvement for analyzing our results. Following the DARTS model [1], [3], we sought input from Summer Institute fellows (SI-fellows) regarding the strengths and weaknesses, as well as differences and similarities across four FIM platforms, and diagnosed usability challenges through structured survey response, across four 4 platforms. We then reviewed best practices across risk communication disciplines and developed a set of interpretive themes against which platforms could be compared. We report results from the diagnosis and assessment stages of the DARTS framework and synthesize our findings into specific targets for the NWPS Emulator redesign. Given the nature of the problem statement, we assumed that usability and information support limitations across the assessed platforms arose primarily from the communication design of the interfaces, as these challenges persisted even among participants with some technical knowledge about FIM platforms.

We analyzed the data collected in two main strands: the qualitative content of assessment of fellow narratives, describing how platform features are interpreted and where they fall short; and the quantitative content of Likert-scale ratings, describing how consistently and how severely those short falls are across platforms. This is necessary because although FIMs could be complex and inherently challenging to interact with, some users could genuinely misinterpret clear messages. Thus, a FIM could demonstrate strong visual design yet could be criticized as challenging. As such, we expected findings from each strand to reinforce the other rather than stand apart to address these potential validity issues.

4.2 Platform Selection and Assessment of FIMs

FIMs selected are made up of two public facing and two credentialed platforms. We identified selected platforms in consultation with team leaders. Given the limited time span of the project, we selected four FIMs, namely NOAA NWPS, FloodID, EarlyFlows, and HurrEvac. These FIM platforms are unique in their operational capacity. The public facing FIM, NOAA NWPS and EarlyFlows, do not require credentials to access them whereas the credentialed FIMs require users to sign up before accessing them. The research team conducted an initial assessment of these platforms to understand (1) its intended audience, public or credentialed; (2) the flood models and forecast types it makes available; and (3) whether it layers in supplementary content such as depth, probability, or scenario-based projections are unique or otherwise before rolling out the survey to SI-fellows.

SI-fellows independently completed a structured assessment of each platform, combining open-ended diagnostic prompts with a Likert-scale rating grid. Fellows are graduate students with varying academic backgrounds including water science, water policies and may already bring prior exposure to the specific platforms under review. No standardized walkthrough was provided beforehand as fellows determined independently how to approach each task, consistent with an interest in preserving real-world user behavior rather than trained performance. The assessment prompts asked fellows to work through representative tasks on each platform and reflect on where the interface helped or hindered their understanding.

4.3 Analysis and Synthesis of FIM Diagnostic

To understand the common communication challenges across the four platforms, we synthesized findings from the research team as well as fellow survey responses. The synthesis was grounded in the collective interpretive experience of the research team. We coded qualitative response in NVivo 15 using a three-tier node structure (A, B, and C nodes), grouped codes into six interpretive themes, and ran matrix coding queries to compare theme frequency across platforms and address the research questions. Coding-reference counts were used as exploratory pattern indicators rather than statistical measures of severity or prevalence. The 1-5 rating-grid items were grouped into eight report-level diagnostic dimensions: visual clarity, navigation/usability, language and terminology, layer/legend clarity, forecast/timing clarity, uncertainty communication, impact communication, and actionability. Rating-grid scores were treated as structured diagnostic indicators rather than statistically representative survey measures as shown in **Figures 2, 5** and **6**.

5. Results

The presentation will be in two folds. First, we will discuss some of the prevalent challenges, strengths, weaknesses, and differences observed throughout the study. This will be supported with screenshots, and quotations from respondents. We will discuss these using the 6 themes as deductively organized in NVIVO. Second, we will discuss key features of our redesigned prototype NWPS emulator developed by the team. Notably, the team was supported by the AWI web developers.

Generally, we find that all four platforms used interactive layers, multiple colors, and symbol-based location markers, but did so in inconsistent ways across platforms. NWPS and EarlyFlows, differed sharply in default landing view. For example, the NWPS platform opened on a national basemap dense with river-gauge symbols, while EarlyFlows opened on a single local extent. FloodID and HurrEvac, required region or storm selection before any flood-relevant layer appeared at all (**Figure 1**). This descriptive baseline sets up the usability and interpretability challenges organized below under the six thematic domains used throughout this study.

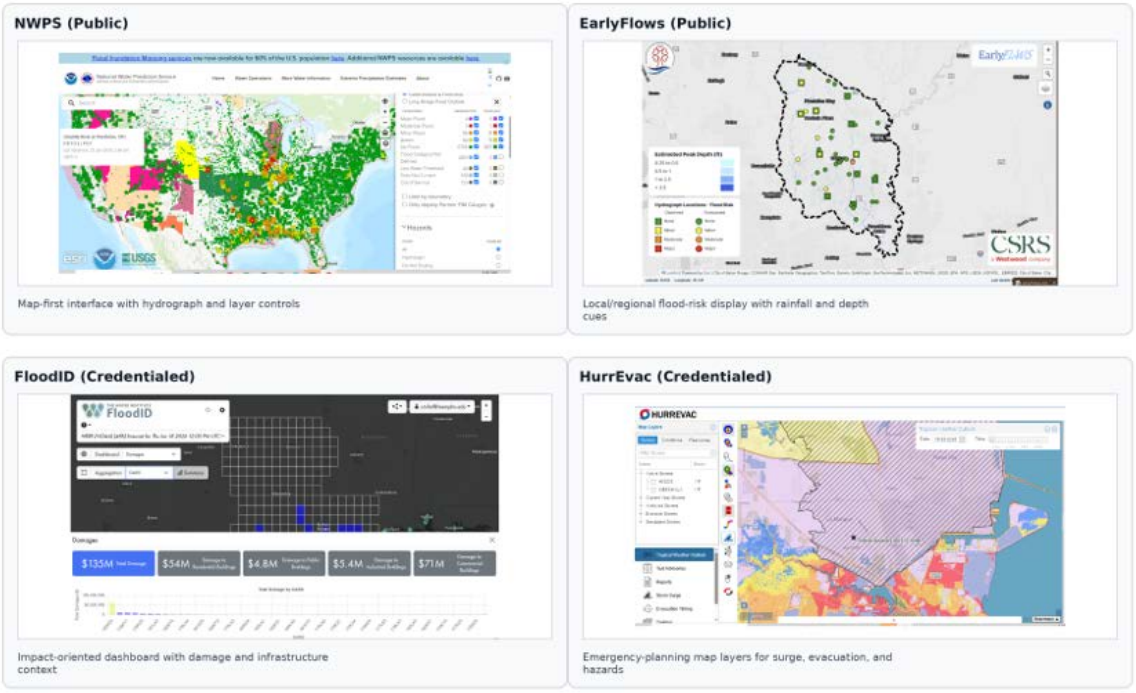


Figure 1. *Representative Screenshot Evidence from the Four Assessed FIM Platforms. The panel Illustrates Differences between Public and Credentialed Interfaces, including Map-first Layouts, Rainfall and Forecast Displays, Impact-oriented Dashboards, and Emergency-planning layers.*

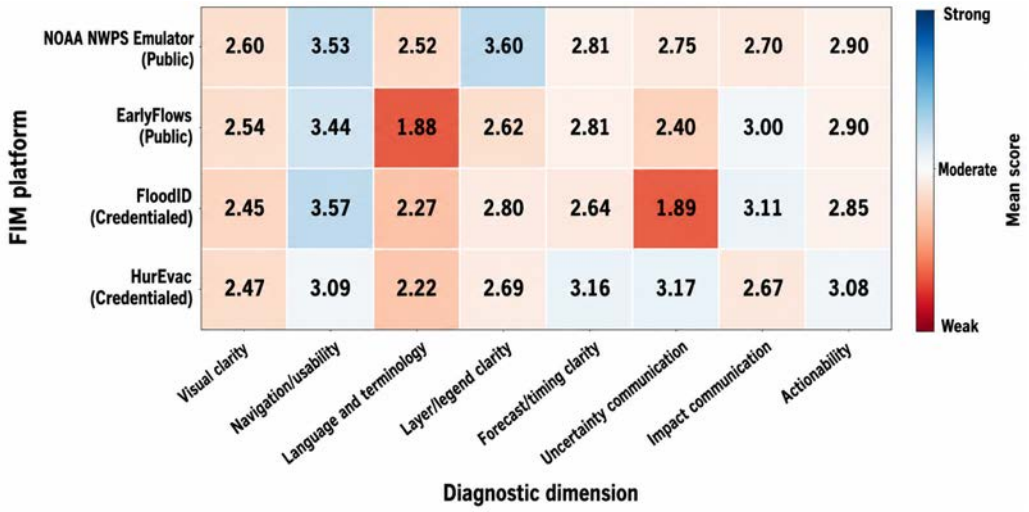


Figure 2. *Likert-scale Ratings by FIM Platform and Diagnostic Dimension. Scores Range from 1 (weak) to 5 (strong) and are Interpreted as Descriptive Diagnostic Patterns.*

The Likert-scale synthesis reinforced the qualitative findings. Navigation/usability received the strongest overall ratings across the four platforms (mean approximately 3.41), while language and terminology was the weakest dimension (approximately 2.22). Visual clarity (approximately 2.52) and uncertainty communication (approximately 2.55) also remained comparatively weak. Platform-level patterns were uneven: the NOAA NWPS Emulator rated highest for layer/legend clarity (3.60),

FloodID rated lowest for uncertainty communication (1.89), and EarlyFlows rated lowest for language and terminology (1.88). These values are used as descriptive evidence to triangulate the qualitative themes rather than as statistically representative platform rankings.

5.1. Usability and User Experience

Notably, NWPS's default view overwhelms first-time users with too many simultaneous layers, while HurrEvac's default view presents minimal visual risk information to register as urgent. Across platforms, reviewers reported that an initial reaction of either overload or emptiness undermined their ability to locate the platform's central message. One respondent described the effect of the dominant "no flood" color coding directly: *"I see almost all of the United States painted green. Only a few areas show signs of red paint. I do not know if the green paints are flood or not"* (NWPS respondent). *"The map itself looks pretty empty — mostly a light blue ocean and a plain-looking land mass. I am not sure if I should be alarmed or not."* (HurrEvac respondent). These differences in visual clusters could be perceived as challenging. In terms of navigation, a reviewer expressed that *"the changing basemap button is a little bit hidden under the gear icon of settings next to the title"* (FloodID Respondent). This visual design consistently surfaced, creating a lack of interface friendliness for responders. This result is consistent with hypothesis 2a.

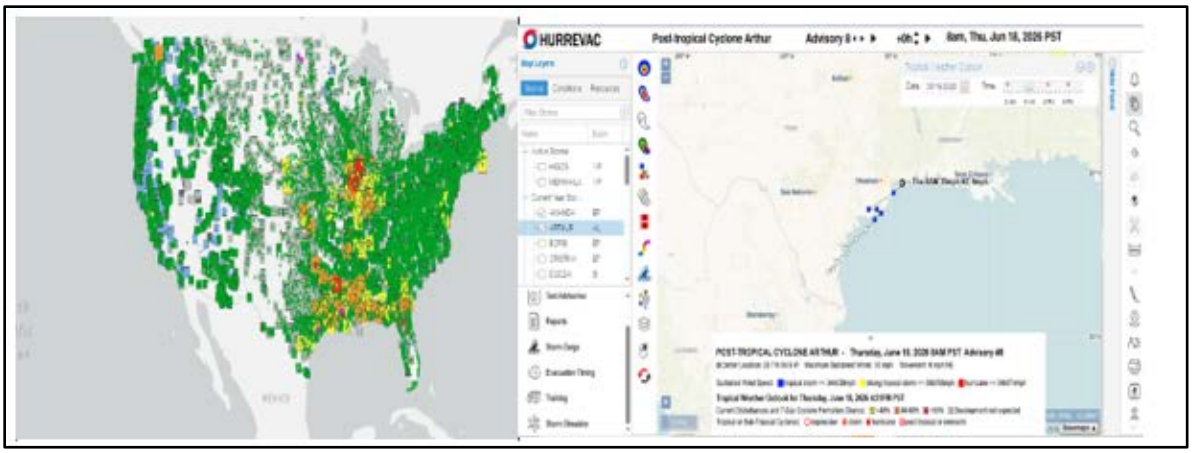


Figure 3. *Different First Interface of NWPS and HurrEvac.*

5.2. Communication of Flood Risk

Across the platforms, a notable observation was the use of symbols and shapes to represent risk information. The analysis revealed that different platforms used different shapes. This was expressed as inconsistent. Thus, reviewers could not reliably distinguish symbol shapes representing different location types. A respondent said, "*Some of the dots are square and others are circles which I am not sure they mean."* (FloodID Respondent). When a platform varying shapes to signal "no risk" is reused elsewhere on the same map, a user must consciously override their first read to check the legend before trusting what they see, slowing exactly the response the platform is meant to enable. This supports H1a and H1b.

Figure 4. *Symbols and Shapes use as NWPS Representation of Risk at Different Legend Interface.*

This form of shape and symbol inconsistencies reveals clear differences across platforms even though; this variation was not consistent based on FIM types per the study. Thus, the use of symbols was merely random across all 4 FIMs.

5.3 Accessibility and Inclusiveness

Technical jargon persisted as a barrier among respondents indicating that this form of barrier originates in the platform's design rather than in the user's unfamiliarity with flood science. This pattern was observed across both FIM types. Reviewers revealed that *"Most aspects of this platform would be confusing for a general-public user... even as someone who is familiar with much of this terminology I am confused by what each layer is intended to show."* (HurrEvac, Respondent). Another reviewer also stated that *"AHPS/NWPS, Hydrograph, Gauge Stage, CFS (Cubic Feet per Second), Exceedance Probability, Inundation, RFC, Return Period."* (NWPS Respondent). This supports H3a: undefined technical language is a barrier to understanding flood risk regardless of a user's background.

5.4 Information Quality

FIMs should be able to achieve information quality through uncertainty and clear precision and scenarios. However, according to our analysis, this was not observed across all platforms. Respondents struggled to determine whether the information displayed represented current conditions or a future forecast. One respondent indicated "*No. They just have short range and long-range terminology which isn't explicit."* (NWPS, Respondent**).** Also, a respondent indicated that "*No. The platform shows forecast-model outputs and risk categories, but no direct probability values, confidence intervals, or likelihood estimates were visible." "The 24-hour observed rainfall totals and described MRMS/gauge data help provide recent observed context and model starting conditions. However, explicit historical flood extent or hindcast flood maps were not visible in the screenshots."* (EarlyFlows, Respondent)**.** Under these uncertainty situations, a user who cannot tell whether flooding has already happened or is still approaching cannot judge how much time remains to act. This can create communication challenges. This result supports hypothesis 2b: platforms with clear flood-extent and uncertainty communication build trust and understanding, and unclear timing actively undermines both.

5.5 Trust and Transparency

Trust and credibility emerged as crucial risk perception factors. Respondents stated *"The NWPS/FIM platform is credible because it is an official NOAA/NWS product, includes source/update information, and separates*

latest analysis from forecast products." (NWPS) Also, it was observed from another respondent stating that *"HURREVAC is strong for forecast timing, forecast products, scenarios, and operational trust because it is built around official NHC/WPC/NWS/NWC-type products and advisories."* (HurrEvac). An important inference is that a user may open a platform, trust it immediately because of its federal branding, and still leave without understanding what it showed them. This partly reinforces H2b by showing that credibility and information clarity are not complementary. FIMs should be designed to achieve both needs.

5.6 Decision Support

Reviewers consistently distinguished between a platform showing flood risk and a platform helping them decide what to do about it, and this gap appeared on both public-facing and credentialed platforms. FloodID was the exception, credited specifically for pairing forecast data with actionable impact information. Evidently, some respondents stated that *"I would add a layer of infrastructure impacts. Also, a simple landing message or banner that says something like 'Here is what is happening right now and here is what you should know'... before I have to dig through layers and panels to figure it out myself."* (EarlyFlows Respondent). Another respondent also stated that *"FloodID includes multiple impact-oriented dashboards, including Roadway Inundation, Search and Rescue, Damages, Gates, and Response Infrastructure. The Search and Rescue dashboard shows depth in feet, which can support emergency-response planning."* (FloodID Respondent). A unique feature observed amongst credentialed FIMs is the actionability feature. This supports H3b: reviewers commended platforms that demonstrated actionability and treated its absence as a shortfall even when the underlying forecast data was accurate.

In summary, the qualitative coding, screenshot evidence, and Likert-scale synthesis converged on the same diagnostic pattern: language and terminology, visual clarity, uncertainty communication, and layer/legend clarity were recurring weaknesses, while impact-oriented information strengthened perceived actionability. **Figure 5** documents how individual rating-grid items were grouped into report-level diagnostic dimensions, and **Figure 6** summarizes descriptive mean scores by dimension. These results provide preliminary support for several diagnostic hypotheses but should not be interpreted as causal or statistically generalizable confirmation.

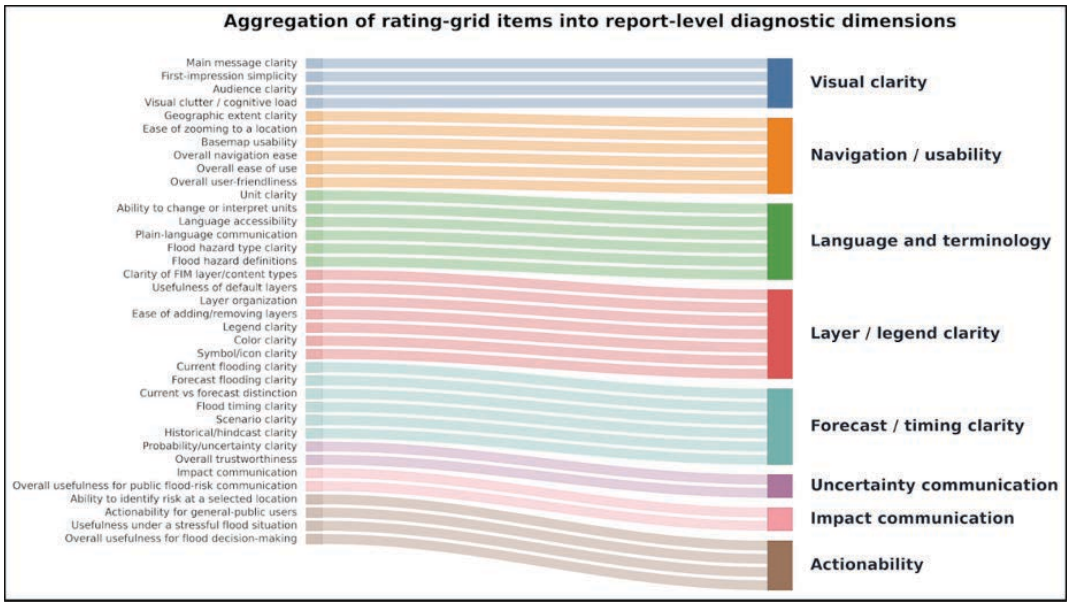


Figure 5. *Aggregation of Rating-grid Items into Report-level Diagnostic Dimensions.*

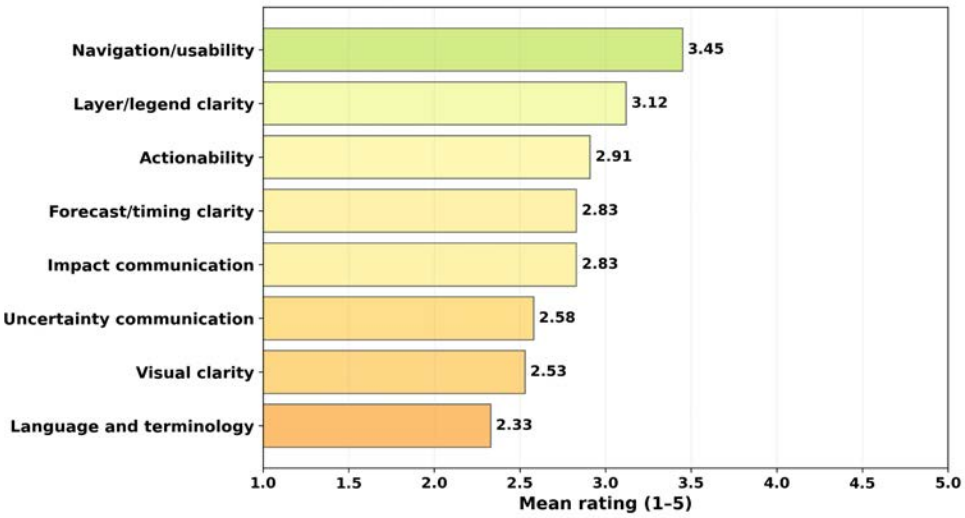


Figure 6. *Updated Mean Rating-grid Scores by Diagnostic Dimension across the Combined Assessment Records. The Figure Summarizes Descriptive Patterns and should be Interpreted Alongside Qualitative Coding and Screenshot Evidence.*

5.7 NWPS Emulator Redesign: Translating Diagnosis into Prototype Features

The second part of the results translated the diagnostic findings into practical changes within the NOAA NWPS Emulator. The assessment indicated that reviewers did not primarily need additional hydrologic information; they needed clearer guidance about what the map showed, whether the displayed information represented current or forecast conditions, and what real-world impacts might require attention. Accordingly, the redesign focused on communication support rather than attempting to create a complete operational decision-support system.

The main redesign feature was an Impact Lens, initially implemented as a Road Impact Lens. The guided interpretation panel explains the purpose of the displayed layers, separates forecast inundation from official warning information, and provides plain-language explanations for technical terms and map products. It also communicates forecast timing and limitations through valid-time, update, and interpretation notes. The Road Impact Lens screens road features within the current map view against NWPS forecast inundation polygons and official warning areas. Because the prototype does not use verified road-closure data, the output deliberately uses cautious wording, including "potentially affected roads" and "roads requiring review," rather than describing roads as closed, unsafe, or impassable. **Figure 7** illustrates the original NWPS interface, the guided Impact Lens, and the Road Impact Lens screening output.

Three near-term design priorities emerged from this translation. First, the opening interface should reduce visual competition by emphasizing active flood information and allowing secondary layers to remain available but off by default. Second, acronyms, flood categories, forecast products, and timing cues should be supported by concise tooltips and guided explanations. Third, impact-oriented summaries should help users move from flood extent to practical screening questions, beginning with roads and later extending to buildings, critical facilities, shelters, and community exposure when validated data are available.

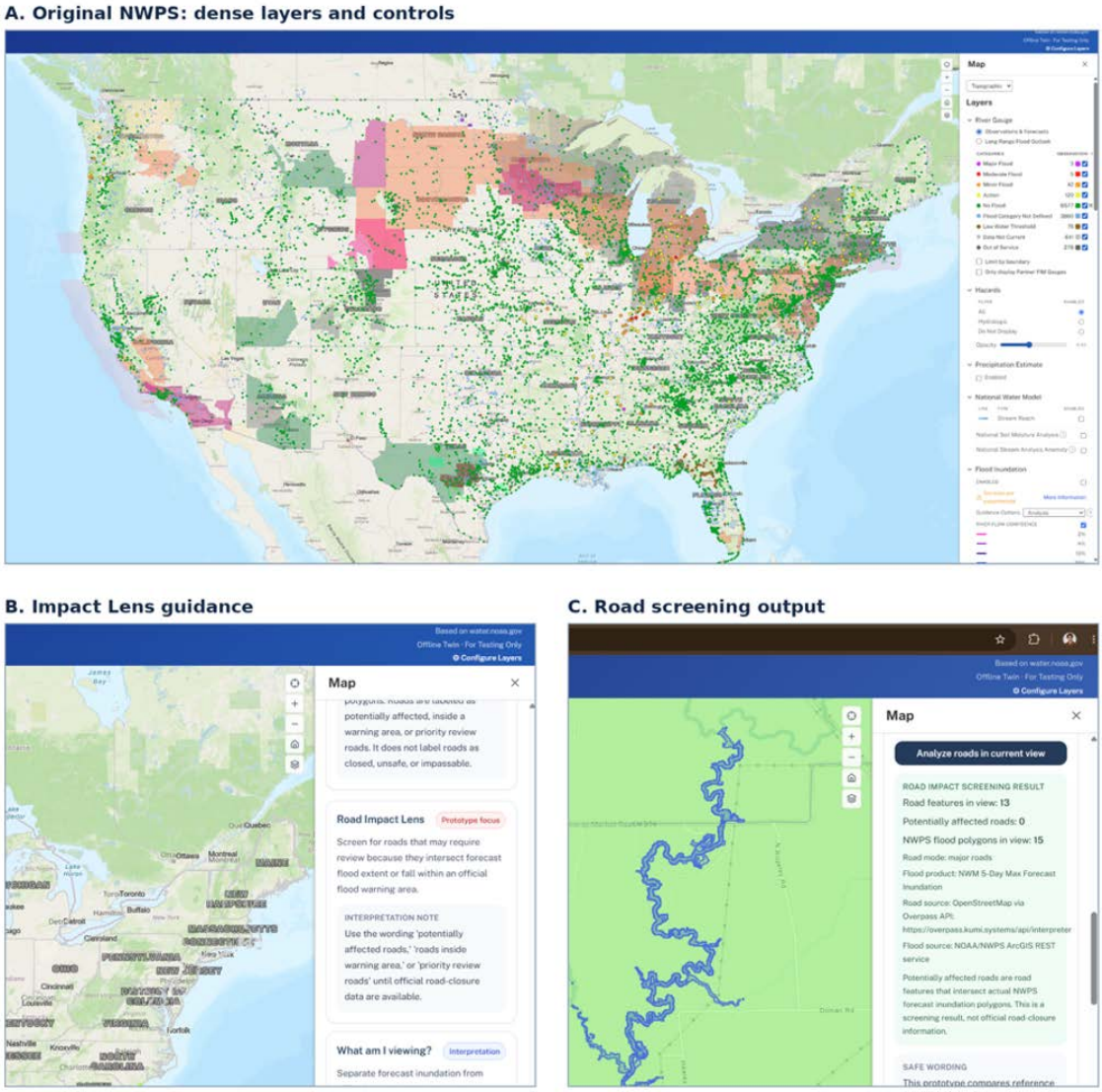


Figure 7. *NWPS Emulator Prototype Features: (a) Original Interface; (b) Impact Lens Guidance; and (c) Road Impact Lens Screening for Roads Potentially Intersecting Forecast Inundation or Warning Areas.*

6. Limitations and Future Work

This study has several limitations. First, the findings are based on an exploratory diagnostic assessment of 13 respondents/assessors rather than a statistically representative sample of public users, emergency managers, or all possible FIM audiences. This poses data saturation concerns. Second, response coverage was uneven: seven fellows assessed NOAA_NWPS, two assessed FloodID, one course coordinator assessed HurrEvac, and three team members assessed all four platforms. Third, the study examined existing interfaces that differed simultaneously in purpose, data, terminology, layers, and design; therefore, it cannot isolate causal effects of individual visualization features. Fourth, the Likert scores and coding-reference counts are descriptive indicators rather than inferential hypothesis tests. Fifth, the study evaluates communication and interface design, not the hydrologic or hydraulic accuracy of the underlying forecast models. Finally, screenshots from credentialed platforms should be treated carefully in public releases because access and permission constraints may apply.

Future work should test the revised NWPS Emulator design with a broader group of users, including emergency managers, hydrologists, and public users with different levels of technical experience. The next DARTS steps would include structured prototype testing, comparison of task completion and interpretation accuracy before and after redesign, and refinement of Impact Lens features using validated datasets. Future development could also extend the Road Impact Lens to critical facilities, buildings, shelters, neighborhoods, and local guidance links where appropriate data and permissions are available.

7. Conclusion

This report builds on Kandel et al. [1] by applying a diagnostic framework to four selected FIM platforms and translating the findings into NWPS Emulator design targets. The analysis combined screenshot evidence, NVivo qualitative coding, Likert-scale rating-grid synthesis, nine fellow assessments, one course-coordinator response, and three team-member assessments in which each team member reviewed all four FIMs. The main result is that the most persistent barriers were interpretation problems rather than basic navigation problems: unclear terminology, weak first-screen clarity, layer and legend confusion, forecast timing and limitation gaps, and uneven impact translation. The literature-based hypotheses were retained as diagnostic hypotheses and examined through triangulation rather than as causal hypotheses tested through a controlled experiment.

Public and credentialed platforms showed different preliminary patterns, but these comparisons should be treated cautiously because response coverage was uneven across platforms. The strongest finding is not a final ranking of platforms; it is the recurring need for clearer explanation structure and more impact-oriented interpretation. Access type alone did not determine communication quality. The most important factors were interface design, terminology, forecast-context explanation, and the degree to which flood information was translated into decision-relevant impacts.

The findings support NWPS Emulator improvements that prioritize clearer default interpretation, plain-language explanations, forecast-timing and limitation guidance, and impact-aware summaries. The most feasible near-term target is a Road Impact Lens that identifies roads requiring review or potentially intersecting forecast extent. This is not an official closure tool; rather, it is a communication bridge that helps users move from mapped flood extent toward practical, impact-aware interpretation. Future work should analyze the full raw response set, expand platform-balanced testing, and evaluate whether these design changes improve comprehension, reduce interpretation burden, and support more confident flood-risk decision making.

Acknowledgements: We acknowledge Danna Villareal: *University of Arkansas*, Megan Vardaman: *University of New Hampshire* and Nana Oye Djan: *Carnegie Mellon University* for their support and guidance in ideation and logistical needs. We acknowledge our academic advisors, Christopher Koliba: *The University of Kansas* and Dr. Rouzbeh Nazari, for their guidance and support. The authors also acknowledge Manjila Singh and Dinuke Nanayakkara Munasinghe: *Alabama Water Institute (AWI)* for their support with NWPS Emulator. We acknowledge the support of the University of Alabama Institutional Review Board for the approval of our human subjects protocol.

We are grateful for the support of the Alabama Water Institute Staff, including Lanna Nations and Hannah Holcomb–CUAHSI organizers, including Julia Masterman who in diverse ways created the perfect environment for the team to accomplish this project.

This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) under award NA22NWS4320003 from the NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA.

Generative AI, including ChatGPT (July 2026), Claude (2026) was used to assist with selected drafting, language refinement, report organization, figure-caption development, and limited code support for summary outputs. All factual claims, citations, code outputs, analytical interpretations, and conclusions were reviewed and verified by the authors, who remain fully responsible for the final content of this report.

Supplementary Materials: Processed data, diagnostic assessment materials, visualization scripts, NVivo codebook and reproducibility documentation are available through the project HydroShare resource (http://www.hydroshare.org/resource/48f3e77af2ea4af1bccb1a7296ade472) and GitHub repository (https://github.com/NWC-CUAHSI-Summer-Institute/visualize-the-rise-fim-diagnostic).

References

[1] S. Kandel *et al.*, "Comparing Flood Inundation Map Features and Diagnosing Decision Support Design Challenges," *Hydrological Processes,* vol. 40, no. 1, p. e70362, 2026.

[2] F. Pappenberger *et al.*, "Visualizing probabilistic flood forecast information: expert preferences and perceptions of best practice in uncertainty communication," *Hydrological Processes,* vol. 27, no. 1, pp. 132–146, 2013.

[3] A. Joshi, S. Sainjoo, J. Kennedy, and M. A. Kenney, "Testing forecast visualization designs for improving agricultural decision support products," *Environmental Research Letters,* vol. 21, no. 9, p. 094015, 2026.

[4] M. D. Gerst *et al.*, "Using visualization science to improve expert and public understanding of probabilistic temperature and precipitation outlooks," *Weather, Climate, and Society,* vol. 12, no. 1, pp. 117–133, 2020.

[5] S. Seipel and N. J. Lim, "Color map design for visualization in flood risk assessment," *International Journal of Geographical Information Science,* vol. 31, no. 11, pp. 2286–2309, 2017.

[6] R. Soares *et al.*, ""The Great Vermont Floods of 2023": A Risk and Crisis Communication Assessment," *Society & Natural Resources,* pp. 1–23, 2025.

[7] M. Agrawala, W. Li, and F. Berthouzoz, "Design principles for visual communication: How to identify, instantiate, and evaluate domain-specific design principles for creating more effective visualizations," *Communications of the ACM,* vol. 54, no. 4, p. 60.

[8] A. C. Costa, J. Bakker, and G. Plucinska, "How and why it works: The principles and history behind visual communication," *Medical Writing,* vol. 29, no. 1, p. 16, 2020.

[9] B. F. Liu *et al.*, "Is a picture worth a thousand words? The effects of maps and warning messages on how publics respond to disaster information," *Public Relations Review,* vol. 43, no. 3, pp. 493–506, 2017.

[10] R. Soares *et al.*, ""The Great Vermont Floods of 2023": A Risk and Crisis Communication Assessment," *Society & Natural Resources,* vol. 39, no. 2, pp. 309–331, 2026.

[11] R. E. Kasperson *et al.*, "The social amplification of risk: A conceptual framework," *Risk analysis,* vol. 8, no. 2, pp. 177–187, 1988.

[12] O. Renn and D. Levine, "Credibility and trust in risk communication," in *Communicating risks to the public: International perspectives*: Springer, 1991, pp. 175–217.

[13] M. D. Gerst, M. A. Kenney, and E. Read, "Using visualization science to inform the design of environmental decision-support tools—A case study of the US Geological Survey Waterwatch," US Geological Survey, 2328-0328, 2025.
[14] National Weather Services. "National Water Center Products and Services." https://www.weather.gov/owp/operations (accessed 2026).
[15] D. Kyei, W. Donner, A. S. Lomeli, and D. Kyne, "From risk perception to readiness: understanding US wildfire preparedness drivers," *Fire Ecology,* 2026.
[16] M. Hagemeier-Klose and K. Wagner, "Evaluation of flood hazard maps in print and web mapping services as information tools in flood risk communication," *Natural hazards and earth system sciences,* vol. 9, no. 2, pp. 563–574, 2009.
[17] M. K. Lindell and R. W. Perry, "The protective action decision model: Theoretical modifications and additional evidence," *Risk Analysis: An International Journal,* vol. 32, no. 4, pp. 616–632, 2012.
[18] L. E. Taylor *et al.*, "Applying the IDEA model to flood risk communication," *Crisis and Risk Communication,* vol. 1, no. 1-2, pp. 218–240, 2025.
[19] H. Nobanee *et al.*, "A bibliometric analysis of objective and subjective risk," *Risks,* vol. 9, no. 7, p. 128, 2021.

Appendix

Theme Leads

Advancing the Use of Quantitative Precipitation Inputs into the NextGen Framework

---

Mohamed Abdelkader

"I am an Assistant Research Scientist at IIHR—Hydroscience and Engineering, University of Iowa. My scientific research interests concern multiple aspects of streamflow measurement, modeling, forecasting, and estimation using ground-based and satellite remote sensing. My work at the Iowa Flood Center focuses particularly on understanding the genesis of flash floods through field experimentation and modeling, as well as the quantification of hydrologic prediction at a range of temporal and spatial scales."

Humberto Vergara

"I am an Assistant Professor in the Department of Civil and Environmental Engineering at the University of Iowa and a Faculty Research Engineer with IIHR-Hydroscience and Engineering and the Iowa Flood Center. I direct the Advanced Hydrology and Warning Applications (AHWA) Lab, where I develop hydrologic prediction systems and decision-support tools for flood forecasting and early warning. My work spans urban, statewide, national, and international scales, integrating remote sensing, numerical weather prediction, hydrologic modeling, and data-driven approaches."

Innovation in Flood Inundation Mapping for Operational Forecasting

---

Anupal Baruah

"I am a Research Scientist at the Department of Geography and Environment, University of Alabama. My research interests include large scale flood inundation mapping and flood forecasting, uncertainty quantification in Process based models, and the application of Earth Observation and AI in hydrology. My favorite hobby is playing badminton, and listening to music"

Sagy Cohen

"I am a Professor at the University of Alabama, with a Ph.D. from the University of Newcastle, Australia. My research interests focus on flood inundation mapping, flood remote sensing and analysis, and global and continental riverine modeling. This marks my seventh involvement in the Summer Institute, with eight of my graduate students participating to date and nine journal papers published stemming from this collaboration. In my free time, I enjoy sailing, kayaking, hiking, and traveling."

Advancing Machine Learning and Artificial Intelligence Frameworks for Enhanced Hydrologic Prediction

---

Mohammad Ali Javidian

"I am an assistant professor in computer science at Appalachian State University (ASU). My research interests include classical/quantum causal inference and probabilistic graphical models for decision making under uncertainty. My research seeks to develop theoretically sound, reliable, and scalable Causal AI/ML approaches with applications in computer systems, water resource research, and healthcare."

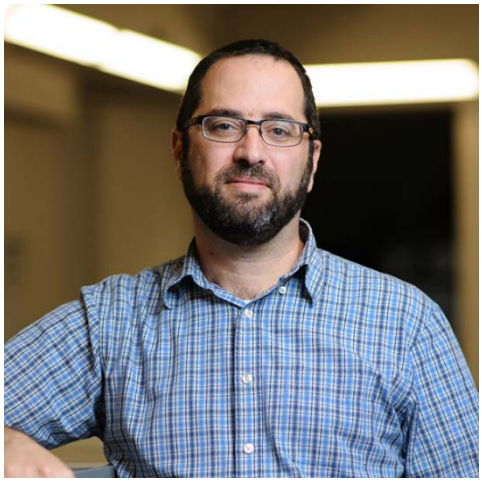
Sushant Mehan

"I work at the interface of water resources engineering and hydroinformatics. My research agenda includes advancing the understanding and physical representation of nutrients' fate and transport to manage water resources using the computing infrastructure. I am also interested in exploring how interactions between climate and land management operations impact water quantity and quality. My passion for working on water resources, especially water quality, is driven by the idea that resolving water problems can solve much unseen unpredicted chaos in the ecosystem. I love to work as a problem-solver on issues related to water sustainability."

Advancing Risk Communication and Public Engagement in Hydrologic Hazard Forecasting

---

Sagy Cohen



"I am a Professor at the University of Alabama, with a Ph.D. from the University of Newcastle, Australia. My research interests focus on flood inundation mapping, flood remote sensing and analysis, and global and continental riverine modeling. This marks my seventh involvement in the Summer Institute, with eight of my graduate students participating to date and nine journal papers published stemming from this collaboration. In my free time, I enjoy sailing, kayaking, hiking, and traveling."

Kelsey McDonough

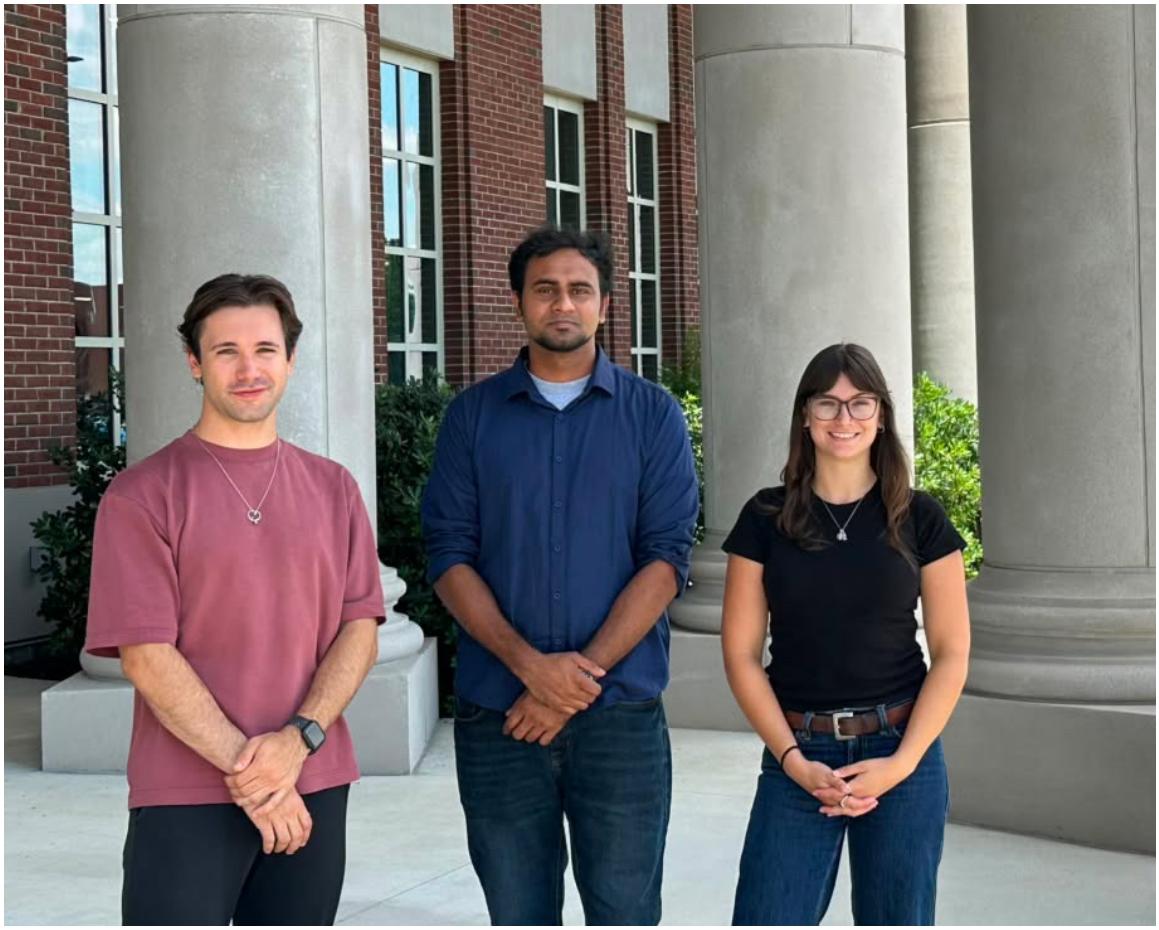
"I am the FloodID Product Lead for The Water Institute, based in Boulder, CO, leading product strategy and growth of our operational flood forecasting technology. This marks my third summer involved with the Summer Institute. My research interests include flood inundation mapping, early flood warning and operational forecasting, and technology transformation for varied applications. In my free time, I love hiking, gardening, and traveling."

Summer Institute Course Coordinators

From left to right: Nana Oye Djan, Megan Vardaman

Summer Institute Teams

Chapter 1: Differentiable Modeling Leverages Sub-Hourly Radar Input for High-Performance Flash Flood Forecasting in NextGen



From left to right: Leo Lonzarich, Azizur Rahman, Jessica Keiser

Chapter 2: Ensemble Streamflow Forecasting for Real-Time Flash Flood Prediction within the NextGen Framework

From left to right: Mohammad Mosavat, Niloufar Soheili, Aldo Andres Tapia Araya, Amirhossein Montazeri

Chapter 3: Sensitivity of Flash Flood Predictions to Precipitation Forcing Uncertainty for Different Spatiotemporal Resolutions

From left to right: Alexander Lastner, Danna Villarreal, Bikas Chandra Gupta

Chapter 4: Evaluating the Sensitivity of HAND Flood Inundation Mapping to River Slope

From left to right: Sebastian Marshall, Zih-Syun Chen, Reza Jamshidi, Pitamber Wagle

Chapter 5: Improving FIM accuracy through stage-dependent Manning's n calibration of HAND-derived SRC

From left to right: Santiago Henao Gomez, Armen Konialian, Alemayehu Dula Shanko, Ali Haider

Chapter 6: Advancing ML & AI Frameworks for Enhanced Hydrologic Prediction

From left to right: Sanjeev Panta, Saddy Rafael Pineda Castellanos, Basit Akinade, Ahmed Omar

Chapter 7: Toward Better Flood Risk Communication: Applying a Diagnostic Framework for Evaluating and Redesigning FIM Visualizations

From left to right: Dominic Kyei, Sohail Ahmed Tufail