# Building a Streamflow Reanalysis Dataset using Deep Learning-Based Geostatistical Signal Separation

Quinn Lee, James Halgren, Taye Akinrele, Sonam Lama, Pratiksha Chaudhari, Josh Cunningham

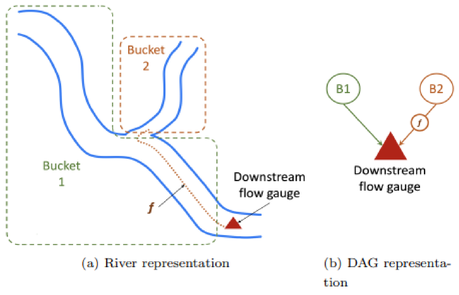
The University of Alabama, Alabama Water Institute

## Abstract

Machine learning (ML) models can be used to predict flows in ungaged basins. If we constrain the ML training with information from observed flows in the same network, we can increase the accuracy of the ungaged flow estimates, We have conducted experiments to show that, with two-basin networks, one downstream and one upstream, an ML model can be trained to provide a higher accuracy estimate of the upstream flows from "ungaged" synthetic basins than can be obtained by training an ML model on individual basins alone.

**To generate a historical streamflow model calibration dataset, machine learning-based streamflow estimates in ungaged basins can be constrained using known downstream flow values as model inputs.**

## Model Structure




(a) River representation (b) DAG representation

Figure 1. Catchment network structure. [1]

## Data Preprocessing

### Catchment pair generation

- Used National Water Model (NWM) Hydrofabric
- CAMELS basins [2] as downstream catchments across continental US
- Upstream catchment selected from the total upstream area (orange) of downstream catchment (pink)

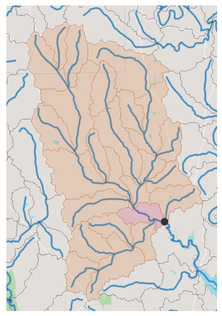


Figure 2. Basin example.

### Data sources

- NWM 3.0 retrospective
- NWM hydrofabric
- CAMELS dataset

### Meteorological forcings (NWM retrospective)

- Wind speed (x & y), air temperature, precipitation, pressure, radiation (shortwave & longwave)

### Static attributes (NWM hydrofabric)

- Contributing area, altitude, stream order, Manning's roughness coefficient, channel geometry, basin geometry, reach length, longitude & latitude

## Site Selection

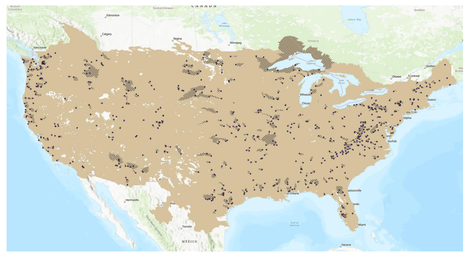


Figure 3. CAMELS gages (black dots) and their upstream areas (crosshatched area).

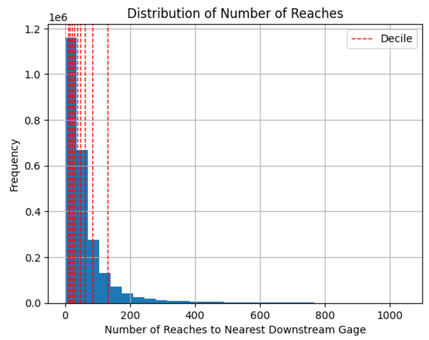



Figure 4. Histogram showing distribution of number of reaches between each NWM reach and its nearest downstream gage.

### Selected Distances

- Let *n* be the number of reaches between a downstream gage and an upstream basin where streamflow is to be predicted.
- Data prepared for *n*=1, 5, 10, 20, 30, 40, 50, 60, 90, 150

Input Datasets

## Model Setup

| Model type | Inputs | Outputs |
|---|---|---|
| Upstream (individual) | Meteorological forcings & catchment geometry of upstream catchment | Streamflow at upstream catchment |
| Combined | Meteorological forcings & catchment geometry of upstream and downstream catchments + streamflow at downstream catchment | Streamflow at upstream catchment |

Table 1: Model setup comparison.

| *n* | Number of preprocessed catchment pairs |
|---|---|
| 1 | 630 |
| 5 | 588 |
| 10 | 520 |
| 20 | 395 |
| 30 | 294 |
| 40 | 213 |
| 50 | 154 |
| 60 | 116 |
| 90 | 54 |
| 150 | 10 |

Table 2: Sizes of input datasets.

## GitHub Repositories

Data Preprocessing

Machine Learning

## Model Results

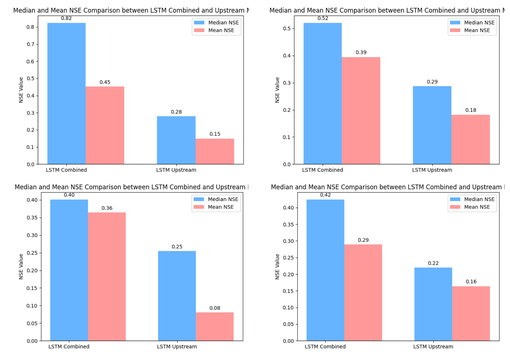



Figure 5. Upstream vs. combined performance comparison summary. *N*=1 (top left), 20 (top right), 50 (bottom left), 90 (bottom right)

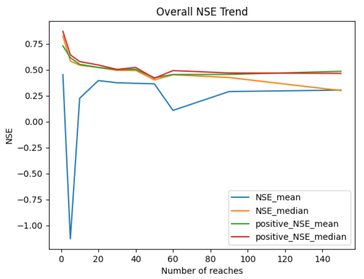



Figure 6. Overall NSE trend across *n* values for the combined model.

| *n* | Average distance between downstream and upstream catchment (km) | Combined model median NSE |
|---|---|---|
| 1 | 1.97 | 0.82 |
| 5 | 9.85 | 0.58 |
| 10 | 19.7 | 0.54 |
| 20 | 39.4 | 0.52 |
| 30 | 59.1 | 0.49 |
| 40 | 78.8 | 0.49 |
| 50 | 98.5 | 0.40 |
| 60 | 118.2 | 0.45 |
| 90 | 177.3 | 0.42 |
| 150 | 295.5 | 0.30 |

Table 2. Median NSEs for the combined model across *n* values.

## References and Acknowledgments

[1] Ramírez Molina, A. A., Frame, J. M., Halgren, J., & Gong, J. (2025). A proof of concept for improving estimates of ungauged basin streamflow via an LSTM-based synthetic network simulation approach. Journal of Geophysical Research: Machine Learning and Computation, 2(2). https://doi.org/10.1029/2024JH000405

[2] Addor, A. Newman, M. Mizukami, and M. P. Clark, 2017. Catchment attributes for large-sample studies. Boulder, CO: UCAR/NCAR. https://doi.org/10.5065/D6G73C3Q

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 authors and do not necessarily reflect the opinions of NOAA. This research utilized the Wukong GPU cluster, the Pantarhei HPC system, the NSF ACCESS allocation EES240087 on Indiana University's Jetstream2, and AWS cloud computing resources 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 and UA HPC Center Services team. Learn more: https://docs.ciroh.org/docs/services/intro