Streamlining Community Modeling through Automated Data Processing
for the Next Generation Water Resources Modeling Framework

AGU25
New Orleans, LA | 15–19 December 2025

Authors

Josh Cunningham[1], Quinn Lee[1], Chad Perry[1], Sonam Lama[1], James Halgren[1], Steven Burian[1], Jordan Laser[2], Mike Johnson[2], AWI Science team members[1] and Open source contributors[3]

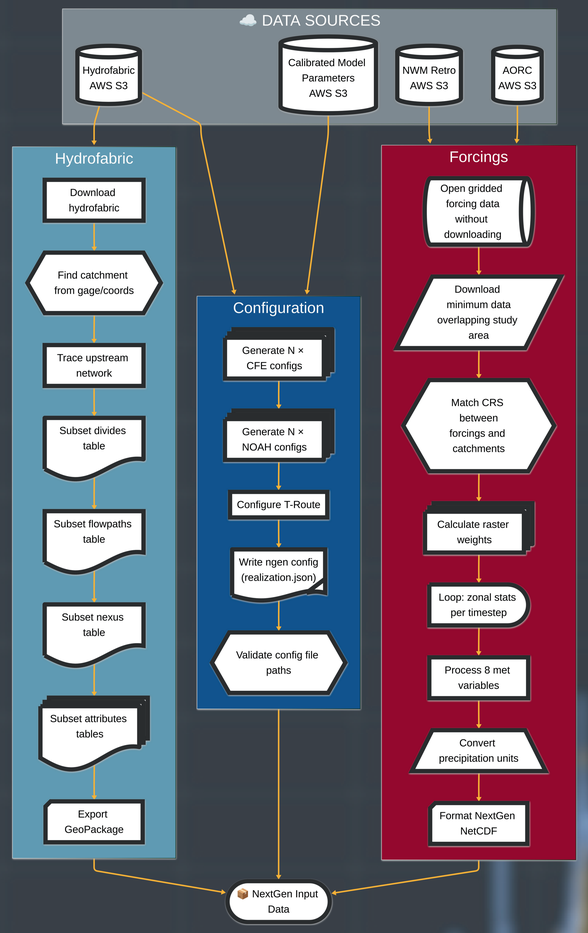
Affiliations

1. Alabama Water Institute, University of Alabama
2. Lynker, Leesburg, VA
3. The National Water Center (NWC), Tuscaloosa AL

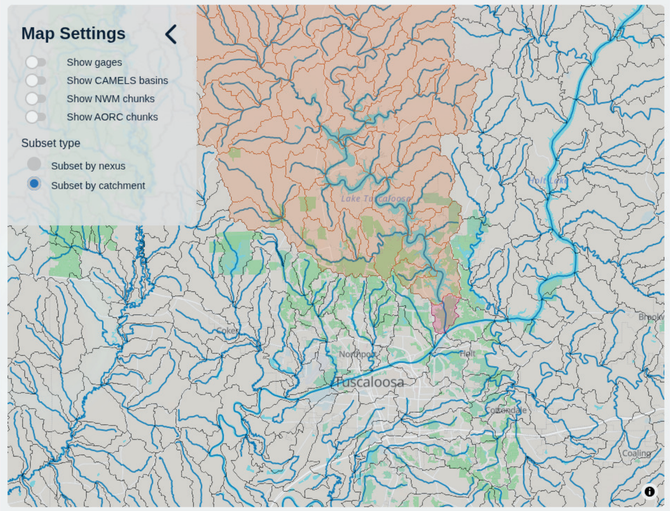
The Challenge

NextGen hydrological simulations require extensive manual data preparation:

- Downloading a national hydrofabric
- Tracing upstream networks
- Extracting database tables
- Processing terabytes of meteorological data
- Computing catchment-averaged forcings
- Generating configuration files.



Select catchments by clicking!



01 Spatial Domain Extraction

Extracts watershed boundaries and attributes from the NOAA OWP v2.2 CONUS hydrofabric using graph-based upstream traversal (igraph).

- Accepts USGS gage ID, lat/lon, or catchment ID
- Extracts 10 related tables: geometry, attributes, network topology
- Outputs self-contained GeoPackage with spatial indices

02 Meteorological Forcing Preparation

Retrieves gridded forcings from cloud-optimized Zarr archives on AWS S3.

- Sources: NWM Retrospective v3 or AORC v1.1 (1km, hourly, 1979–present)
- Parallel S3 downloads via custom S3ParallelFileSystem
- Precise catchment averaging with exactextract
- Multi-core processing with shared memory (~60% memory reduction)
- Local caching prevents redundant downloads

03 Model Configuration Generation

Generates all NextGen configuration files from hydrofabric attributes and calibrated parameters (when available).

- Per-catchment CFE and NOAH-OWP-Modular configs
- T-Route routing with memory-aware loop sizing
- Complete realization.json assembly

```
22:06|josh@aradon:~$ time uvx ngiab-prep -i gage-10154200 --start 1980-10-01 --end 2020-10-01 --source aorc -o poster_speed -sfr
2025-12-10 21:58:56,562 - INFO - Found cat-2863631 from gage-10154200
2025-12-10 21:58:56,562 - INFO - Processing cat-2863631 in /home/josh/ngiab_preprocess_output/poster_speed
2025-12-10 21:58:56,917 - INFO - Subsetting hydrofabric
2025-12-10 21:58:57,877 - INFO - Subsetting complete.
2025-12-10 21:58:57,877 - INFO - Downloading forcings from 1980-10-01 00:00:00 to 2020-10-01 00:00:00...
2025-12-10 22:03:16,065 - INFO - Successfully cached gridded forcing data
2025-12-10 22:03:16,784 - INFO - Computing zonal stats in parallel for all timesteps
2025-12-10 22:03:18,677 - INFO - Total steps: 8, Number of time chunks: 1, Number of variables: 8
Forcings processed in 26.585591 seconds ━━━━━━━━━━━━━━━━━━━━━━━━ 100% 8/8 •  Elapsed Time: 0:00:23  Remaining Time: 0:00:00
2025-12-10 22:03:45,264 - INFO - Forcing generation complete!
2025-12-10 22:03:48,718 - INFO - Creating configuration from 1980-10-01 00:00:00 to 2020-10-01 00:00:00...
2025-12-10 22:03:49,031 - INFO - downloaded calibrated parameters for gage-10154200
2025-12-10 22:03:49,052 - INFO - Realization creation complete.
2025-12-10 22:03:49,052 - INFO - All operations completed successfully.
2025-12-10 22:03:49,052 - INFO - Output folder: file:////home/josh/ngiab_preprocess_output/poster_speed

real    4m54.908s
```

Performance Features

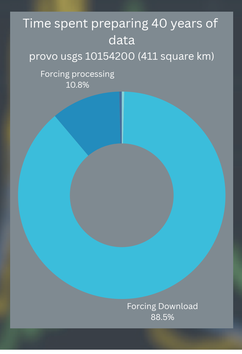
- Parallel S3 range requests reduce data transfer time
- Shared memory blocks allow fast multiprocessing of forcings
- Memory-aware chunking automatically adapts to available system RAM
- Pre-built network graphs enable sub-second upstream traversal even for watersheds with thousands of catchments

Our Solution

- Open-source Python tool automating the complete data preparation pipeline
- A single command transforms location identifier into ready-to-run NextGen simulation package
- Available via web map interface or CLI—no local installation required with uvx.

Community Impact

- Reduces data preparation from hours of manual work to minutes of automated processing
- Reproducible CLI workflows can be shared, scripted, and version-controlled
- Cloud-native data access eliminates need for local terabyte-scale storage
- Web map interface lowers barrier for users unfamiliar with command-line tools
- Extensible architecture accommodates new model configurations as NextGen evolves
- Open source and community maintained—contributions welcome



Future work

- Expand configuration generation to more models, currently only supports cfe+noah-owp-modular and lstm.
- Refactor submodules for easier use in external scripts
- Preprocess 40 year retrospective forcings and save the output in s3 (should be roughly 3% the size of current gridded forcings
- Update to fetch forecast forcing data

Funding & Acknowledgements

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.

This research utilized AWS resources managed by CIROH Cyberinfrastructure. The authors appreciate support from the CIROH Cyberinfrastructure team. Learn more: https://docs.ciroh.org/docs/services/intro