Creating a Community Dataset for High-Speed National Water Model Data Access

Sepehr Karimi[1], James Halgren[1], Joshua Cunningham[1], Karnesh Jain[1] , Arpita Patel[1], Jordan Laser[2], Matt Denno[3], Sam Lamont[3], Benjamin Lee[1], Irene Garousi-Nejad[4], Anthony Castronova[4], Rohan Sunkarapalli[1], Manjiri Gunaji[1], and Steven Burian[1]

[1] The University of Alabama, Tuscaloosa, AL [2] Lynker, Leesburg, VA [3] RTI, Fort Collins, CA [4] CUAHSI, Arlington, MA

Abstract

The National Water Model (NWM) input and output datasets, made freely available on cloud platforms through the NOAA Open Data Dissemination program, represent a treasure of hydrologic information to support research and accelerate improvements in continental-scale hydrologic prediction. The existing format of the NWM files reflects a reasonably efficient configuration for operational simulations, but the cloud-stored files become cumbersome when used for case studies and other retrospective analyses common in research scenarios. CIROH has developed a community-focused dataset to accompany the NWM files that, when combined with simple methods of parallel computing, allows flexible, high-speed access to the NWM operational and retrospective datasets. While this auxiliary reference dataset will be of great value to the hydrologic research community, the methods may also be a guide for provisioning output datasets in future operational hydrology modeling programs. We present here the details of the development of this new dataset along with examples of applications in model evaluation and input preparation for retrospective simulations.

Introduction

The National Water Model (NWM) is a water forecasting model that simulates streamflow in the continental United States, Hawaii, and Puerto Rico.

•NWM OUTPUT

- 40+years retrospective dataset (v2.1:1979 – 2020, v3.0:1979 – 2023)
- Operational dataset since 2018 updated daily

The NWM output data is stored in NetCDF

Challenges with the native NetCDF

- **File Size**: high disk space use, 1 TB+ for operational data, 100 TB+ for the entire retrospective dataset.
- **Complexity**: Computationally expensive.

Methodology

Generating Zarr Files with Kerchunk

What We Did:

- Utilized the Kerchunk library to create Zarr files.
- Generated datasets for both Operational and Retrospective NWM output.
- Made it publicly available on Amazon S3 bucket.

Why:

- Facilitates efficient data storage and retrieval.
- Enables efficient comparative analysis and evaluations.
- Provide pathways for forcing data preparation for NextGen simulation

Benchmarking Scenarios

Comparing the data use and data retrieval performance between Zarr and NetCDF

Output:

- Retrospective 1 year
- Short range 18 hours
- Medium range 240 hours

Environment:

- Cloud – 16 core CPU
- Local – 16 core CPU

Run method:

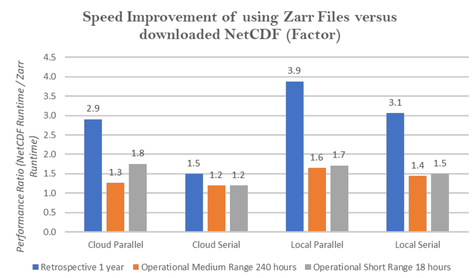
- Parallel
- Serial

Results

Data retrieval performance

Compute Resource / Data Source/Access Pattern	Cloud				Local			
	Zarr Parallel	NC Parallel	Zarr Serial	NC Serial	Zarr Parallel	NC Parallel	Zarr Serial	NC Serial
Retrospective 1 year	12 m 30 s	36 m 18 s	3 h 17 m	4 h 55 m	37 m	2h 22 m	6 h 19 m	19 h 26 m
Operational Medium Range 240 hours	18.2 s	23 s	2 m 23 s	2 m 51 s	28 s	46 s	3 m 7 s	4 m 30 s
Operational Short Range 18 hours	4 s	7 s	10 s	11.9 s	5.5 s	9.4 s	13.8 s	20.6 s

- Zarr headers outperform Native NetCDF in every single scenario
- Zarr performs up to 3.9 faster compared to NetCDF

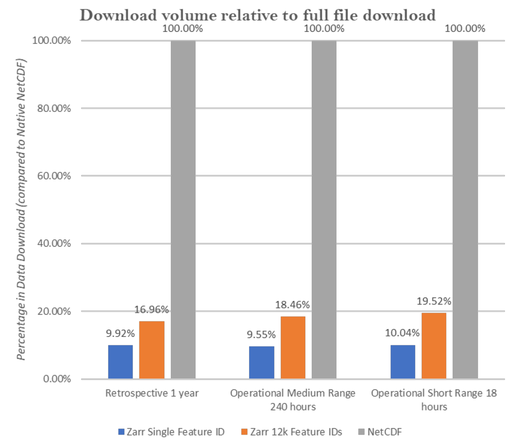


Data use comparison

Comparing Data usage: Zarr vs Native NetCDF

	Zarr 1 Feature ID	Zarr 12k Feature IDs	Native NetCDF
Retrospective 1 year	39.2 GB	78.7 GB	395.3 GB
Operational Medium Range 240 hours	318.7 MB	617.3 MB	3336 MB
Operational Short Range 18 hours	24.9 MB	48.4 MB	248 MB

- Zarr use significantly less data compared to NetCDF
- Data use is reduced up to 90% with Zarr headers



Data Access

dataaccess.ciroh.org

Scan the QR code to access our products:

- NWM Retrospective Zarr Dataset
- NWM Operational Zarr Dataset
- Interactive data download instruction Jupyter notebooks

Acknowledgements

Funding for this project was provided by the National Oceanic & Atmospheric Administration (NOAA), awarded to the Cooperative Institute for Research to Operations in Hydrology (**CIROH**) through the NOAA Cooperative Agreement with The University of Alabama (NA22NWS4320003)