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

Presenting Author: Sepehr Karimi, PhD
mkarimiziarani@ua.edu

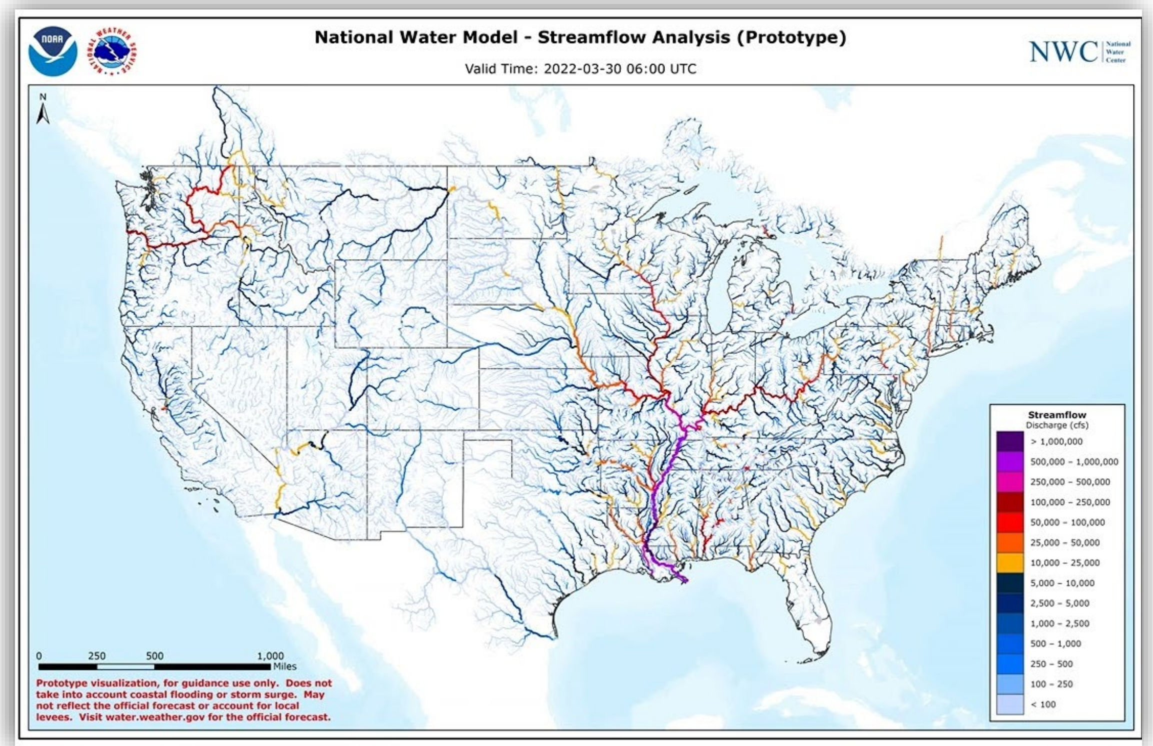
Sepehr Karimi, James Halgren, Arpita Patel, Karnesh Jain, Jordan Laser, Matt Denno, Sam Lamont, Benjamin Lee, Irene Garousi-Nejad, Anthony Castronova, Rohan Sunkarapalli, Manjiri Gunaji, and Steven Burian

January 31st, 2024
AMS Annual Meeting

Baltimore, MD

National Water Model Output Data

- The National Water Model (NWM) is a water forecasting model that simulates streamflow in the continental United States, Hawaii, and Puerto Rico.
- NWM dataset
    - 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 format
    - **File Size**: high disk space use, 1 TB+ for operational data, 100 TB+ for the entire retrospective dataset.
    - **Complexity**: Computationally expensive.



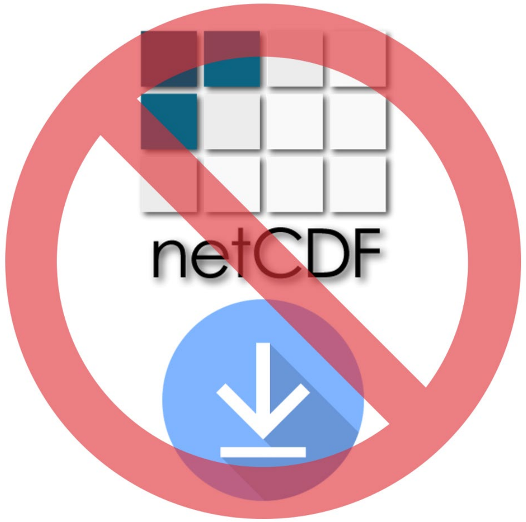
National Water Model - Streamflow Analysis (Prototype)

Valid Time: 2022-03-30 06:00 UTC

Prototype visualization, for guidance use only. Does not take into account coastal flooding or storm surge. May not reflect the official forecast or account for local levees. Visit water.weather.gov for the official forecast.

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



Zarr

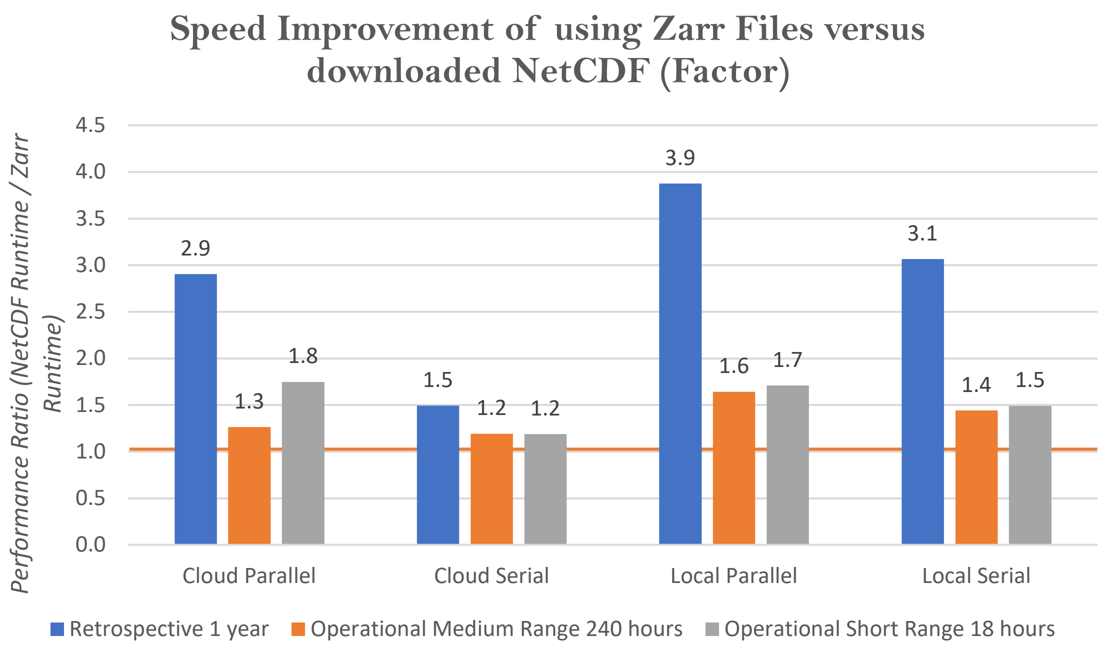
Benchmarking

- 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

Data retrieval performance

Compute Resource	*Cloud*				*Local*			
Data Source/Access Pattern	*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

- *Efficient Runtimes:*
  - Zarr runtimes outshine Native NetCDF in all different scenarios.

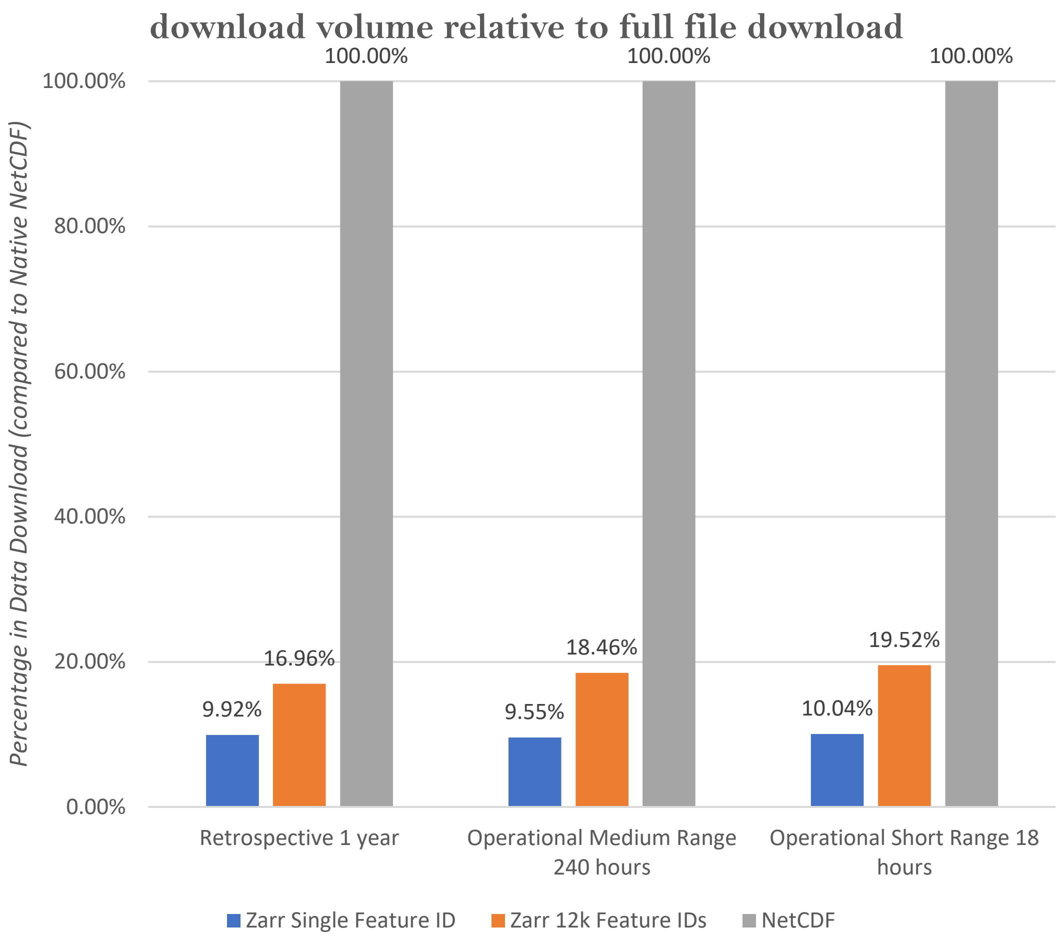


Speed Improvement of using Zarr Files versus downloaded NetCDF (Factor)

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



Dataaccess.ciroh.org

- Retrospective Zarr Dataset
- Operational Zarr Dataset
- Interactive data download instructions – Jupyter notebook

dataaccess.ciroh.org

Thank you!

Sepehr Karimi
mkarimiziarani@ua.edu

dataaccess.ciroh.org