Table of Contents

Project Oversight

PI/CPI – Exe Team	N/A
Primary Mentor	Rajesh Kalyanam
Co-Mentor	Anthony Castranova (CUAHSI)
Technical Advisor	Irene Garousi-Nejad (CUAHSI)

Project Description

Provide a brief summary of the project.

Hydrologic modeling both requires the use of high-resolution retrospective micro-meteorological data and produces high-resolution outputs comprising 10s of variables for the area of interest. For long-term simulations (5 years or more) of the National Water Model's WRF-Hydro, these datasets can quickly run into the terabytes. These raw datasets are in the NetCDF format which presents challenges for extracting specific variables and subsetting to the region and period of interest. This project would involve evaluating various cloud-optimized data formats such as Zarr or Parquet, benchmarking various chunking/subsetting operations, and identifying the best format to support these operations on a variety of retrospective datasets. The transformed (cloud-optimized) data would then be stored in OSN (Open Storage Network) for public access via popular Python packages such as XArray or Dask.

In Scope

Identify the work that will be needed to deliver the project and the key deliverables of that work. For a phased project, identify the work and key deliverables by phase. As needed, describe the deliverables to provide clarity.

The project involves the following components:

1- Get familiarized with the AORC, retrospective, and WRF-Hydro output data currently hosted on the Anvil HPC system to identify the data and dimensional (spatial/temporal) variables.

2- Evaluate various methods of transforming NetCDF4 data into cloud-optimized formats including the optimal chunking configuration.

3- Process various datasets such as retrospective data, WRF-Hydro workflow outputs, and forcing data and store on the Open Storage Network (OSN).

4- Develop example Jupyter notebooks that demonstrate how to access the data from OSN, subset, aggregate, and visualize it.

5- Provide documentation for processing and transferring other cloud-optimized datasets to OSN.

Key Deliverables	Description
Python code for accessing and summarizing the various datasets on Anvil	This involves reading chunks of the data at a time and producing summary metadata of the dataset for use in determining optimal chunk sizes
Python code for chunking and transforming the raw NetCDF4 datasets into a cloud-optimized dataset	This involves evaluating various methods for chunking/cloud-optimization including Zarr, kerchunk, parquet, etc. This would also involve running HPC jobs efficiently to transform the entire datasets once an appropriate method/data format has been determined
Hosting cloud-optimized data on OSN	This involves transferring the resulting data to OSN via S3 commands and ensuring that appropriate metadata and access controls have been set up on OSN for public access
Documentation and user training	This involves developing Jupyter notebooks that demonstrate how to access the data from OSN and carry out various operations on the dataset. This also includes developing documentation for the code that was used to process and optimize the dataset and transfer to OSN

Out of Scope

Identify and describe items that could be assumed to be in the project scope but are not.

We will not optimize for a particular science use case for I-GUIDE. This project is intended to be a generalized exploration of efficient cloud-optimization of a large NetCDF dataset. The assumption would be that once stored in a cloud-optimized format, any of the data variables, or subsetting configurations can be applied to the data.

Desired Outcome

Describe the expectations of the future state after the project has been delivered. Include new capabilities that will be possible because of the delivered project outcome.

When the project is delivered, we will have guiding principles and code examples for how to cloud optimize version the kinds of NetCDF4 datasets relevant to the National Water Model. We will also have these example cloud-optimized datasets hosted on OSN along with instructions on how to access this data and subset for various temporal/spatial domains of interest.

Users of the I-GUIDE platform will be able to access this dataset from a Jupyter notebook that demonstrates how to subset, aggregate, and visualize for their domain of interest.

Other users will also be able to access the data directly from OSN using the appropriate Python libraries.

The cloud-optimization of WRF-Hydro outputs will be integrated into the post-processing for WRF-Hydro to enable efficient retrieval of specific variables of interest.

Project Team Engagement

Describes the overall structure and organization of the project team, as known at this time, and the relationship of roles and/or teams. Indicate any known issues with availability of the project team resources that will be needed.

NAME	TITLE	RCAC TEAM	PROJECT ROLE

Appendix

As needed, include any supplemental information that will elaborate or provide examples for any of the aforementioned sections in the Project Charter.