Parallel and Distributed Data Visualization on the Poseidon Cluster at WHOI

Overview, Goals and Prerequisites:

This workshop assumes knowledge of some basic programming concepts including variable declaration, boolean operators, loops, lists, dictionaries, conditionals and functions. This workshop will use **Python**. If you need to brush up on any of these concepts in Python, the WHOI Python Carpentries workshop website is a good place to start.

The goal of this workshop will be to provide attendees with an introduction to plotting and processing geophysical data stored in tabular (e.g. CSVs) or hierarchical (e.g. NetCDFs) formats.

Why Python?

Based on our experience, Python has the most mature and accessible tools for big data exploration and visualization. Other languages such as MATLAB, Julia, and R offer similar tools, but fall short in many areas. Additionally, many scientific analysis codes are now being written exclusively in Python. So, understanding the basic functions that make up these codes will be worthwhile!

Pre-workshop to-do

- Set up a Conda environment and make sure you can run a Jupyter notebook (instructions will be provided in the following GitHub Repository: https://github.com/anthony-meza/WHOI-PO-HPC)

Workshop Schedule

2:30 PM – 4:30 PM on Thursday October 3rd, Clark 331

Pizza will be provided after the workshop!

Part I (45 minutes): Python, Pandas, Dask and Xarray (lecture/discussion-style)

I. Why Python? (And not MATLAB/Julia)
II. The Big Data “Problem”
III. Introduction to Parallel and Chunking
IV. Efficient Data Structures (CSV vs. Parquet)
V. Introduction to Xarray and Dask

Break (15 minutes)

Part II (1 hour): Data Analysis and Visualization on Poseidon (live-coding)

I. Managing Miniconda and Terminology
    1 slide: Custom Miniconda Installation on Poseidon and using Miniconda to manage environments
    1 slide: HPC terminology (core, node, worker, cpu, scheduler, partition, memory)
II. Running a Jupyter notebook on Poseidon (live-coding)
III. Example: Subsetting and saving ERA5 data using Xarray
IV. Example: Visualization of ERA5 data
V. Example: Temporal Averages (Simple Averages, Climatologies, Trends)
VI. Example: Spatial Averages (Simple Averages, Regional Averages)
VII. (Optional) Example: Filtering (Low-pass filters, high-pass filters)
VIII. Additional Tips and Best Practices
    1 slide: Tmux/screen
    1 slide: Requesting HPC resources (memory, partition, nodes, etc.)