STOFS Data Subsetting Tool

Abstract

Developing a subsetting tool to access ocean forecast data is critical and addresses an immediate need for maximizing the utilization of STOFS model output. STOFS-2D global stands out as one of the most refined and high-resolution global ocean forecast datasets available. This project aims to enhance the functionality of xarray-subset-grid, which is a library that efficiently implements subsetting operations for datasets of various formats (H5netCDF, Zarr, NetCDF4). It includes -

1. Evaluating performance of the library against STOFS data and creating proper documentation for it.
2. A service to run subsetting jobs on demand (through email / API)
3. Choosing efficient methods for subsetting (including structured and unstructured grids), which may involve using existing packages like thalassa, shapely, XUGrid, etc.

Why this project?

I'm eager to contribute to the xarray-subset-grid project with IOOS for several reasons. Firstly, I'm deeply passionate about data analysis, visualization, and climate studies. This project presents a unique opportunity to combine my interests and skills while making meaningful contributions to the field .This project aims to improve data accessibility by efficiently subsetting STOFS and oceanographic data, benefiting users with restricted internet access or compute resources. It empowers researchers and enthusiasts to access essential ocean forecast data for analysis and visualization, irrespective of their technological constraints. Additionally, IOOS's mission resonates with me as it aligns with my values of environmental stewardship and leveraging technology for the greater good. By participating in GSoC with IOOS, I aim to deepen my Python expertise, gain hands-on experience with oceanographic data, and play a part in enhancing accessibility to crucial ocean forecast data.

Technical Details

Libraries involved

1. **Xarray** : Handling labeled multi-dimensional arrays and datasets for efficient data manipulation.
2. **Dask** : Utilizing parallel computing for scalable processing of large datasets.
3. **Zarr** : Storing chunked, compressed, and high-dimensional arrays.
4. **Fsspec / S3FS** : A unified interface for accessing filesystems and cloud storage.
5. **Kerchunk** : Employing efficient chunking of data.
6. **Thalassa** : an established package in the ocean modeling community.
7. **Xugrid** : Providing tools for working with unstructured grid data.

Languages and Technologies -

- Python: Primary language for development, data manipulation, and algorithm implementation.
- Cloud Storage and Resources - Utilizing cloud services for scalability and accessibility of data processing tasks.
- Email Integration: Implementing a service to run subsetting jobs on demand via email.
- Portal / API - Flask / FastAPI , XPublish

References

HDF5 and Zarr Performance Comparison

Kerchunk Enhancements (Issue #42)

Kerchunk Documentation

NOAA STOFS 2D Global

NOAA STOFS 3D Atlantic

NetCDF CF Metadata Conventions

UGrid Conventions

cf-pandas , cf-xarray

IOOS Compliance Checker

Mentor Conversations

Subsetting tool · Issue #57 · ioos/gsoc (github.com)

Based on the conversation and my initial testing of the sample codes, , here are some key things I've learned:

- The project's current focus on the STOFS-2D-Global model output and the plans to expand to other model output types in the future.
- The performance differences between running the code on CPU versus GPU hardware, and the potential need to optimize for different hardware configurations.
- The potential migration from HDF5/netCDF4 to the Zarr format, although the consensus on the faster format is divided.
- The future plans to implement a subsetting service and potentially include plotting capabilities.
- The mentors acknowledged the errors I encountered while trying to subset the dataset using the xarray-subset-grid package and encouraged me to document them for inclusion in my proposal.

Development Experience

I have a Github profile where I host my code and contributions. Here are a few highlights of my development experience -

1. Open-Source Contributions - I have used a lot of Open-Source libraries and packages, and host all my code on my Github profile. I've contributed to Open-Source through hackathons and challenges at my university. This will be my first time contributing to a major Open-Source organization and I intend to work hard towards it.
2. Personal Projects - I have worked on various coding projects, including web scraping scripts, Wi-Fi Scanner/Deauther, chrome extensions, and a question-answer chatbot using Python and Llama. These projects demonstrate my skills in Python programming, data manipulation, and application development.
3. Courses - My relevant academic coursework includes Data Structures, Design and Analysis of Algorithms, Probability and Statistics, Object Oriented Programming, Web Technologies, Introduction to Machine Learning, and Database Management Systems.

Schedule of Deliverables

May 1 - May 26 — Community Bonding Period

Getting familiar with the team members and mentors at IOOS, discussing the project in detail. Understanding the code and requirements from the mentors' perspective.

I also intend on participating in the 2024 IOOS Code Sprint virtually, and joining the community mailing list and the slack channel.

I propose a timeline of 12-16 weeks, encompassing 350 hours and an average of 3-4 hours of work per day.

May 27 - July 11 — Phase 1

1. Determining efficient methods of subsetting data
   This includes conducting benchmarking and comparative analysis to identify the most efficient subsetting methods.
2. Evaluating Performance of Xarray-subset-grid
   Assess the performance of the current xarray-subset-grid implementation against NOAA/NODD STOFS datasets.
3. Developing demonstration notebooks for STOFS and other oceanographic data

Creating documentation to illustrate the usage of xarray-subset-grid for accessing and subsetting STOFS and other ocean data.

4. Integrating xarray-subset-grid with existing packages

   Integrate existing packages in use by the ocean modeling community like Thalassa, Xugrid, etc.
5. Unit Testing and Test Driven Development

   Validation of code with python frameworks like pytest or unittest, and Continuous Integration with Github Actions

July 12 - Aug 19 — Phase 2

1. Enhancing code with zarr, dask, kerchunk for efficiency

   Implement optimizations like using Zarr virtual storage, Dask parallelization to speed up subsetting jobs, kerchunk to improve efficiency and scalability.
2. Expanding the current framework of STOFS-2D-Global to include other formats.
3. Ensure the library is model-agnostic

   Expand the functionality of xarray-subset-grid to work with other public datasets.
4. Refactoring Code

   Minimizing dependencies, ensuring an efficient and lean implementation, adding documentation for the workflow of code and enforcing CF Conventions.
5. Subsetting job service, that operates via email

   Leveraging cloud resources to develop a service or an API for subsetting. (Flask / FastAPI, XPublish)

Aug 19 - Aug 26 — Final Week

Complete any remaining tasks, ensure all code is thoroughly tested, and publish demonstration notebooks to the xarray-subset-grid repository and IOOS CodeLab.

Sep 3 - Nov 4 — Extended Timeline

1. Migrating new datasets to zarr format (or virtual zarr)

   Store new datasets as Zarr format and access old datasets with kerchunk and virtual Zarr files.
2. Any other extended deliverables or tasks discovered throughout the timeline.
3. Creating a portal to view, subset and download STOFS data.

   A basic portal with tools to define a region (either manually or polygon drawing on a map) then running #3 Job service.

Additional Background

- What are your goals for participating in GSoC in relation to your career or future studies?

  My goal for participating in GSoC with IOOS is to advance my career in data analysis and visualization while contributing to climate studies. By enhancing xarray-subset-grid, I aim to deepen my Python skills and gain experience with oceanographic data. This opportunity aligns perfectly with my interests and aspirations, allowing me to make meaningful contributions to a field I'm passionate about.

- What is your preferred method to communicate with your project mentors during the coding period (virtual check-in meetings, Slack, GitHub issue, email)?

  The preferred method of communication is through Github Issues or email. We could have virtual check-in meetings every week to update them about my weekly progress , or this could be done for each phase as well. Given the time difference between PST/EST and IST, I propose the below timings for the virtual meetings, as per the mentors' availability.

IST	PST	EST
05:00 to 07:00	19:30 to 21:30 (previous day)	16:30 to 18:30 (previous day)
19:00 to 21:00	06:30 to 08:30 (same day)	09:30 to 11:30 (same day)
22:00 to 01:00	09:30 to 12:30 (same day)	12:30 to 15:30 (same day)

- Is there anything your mentors should know about your work schedule or studies during GSoC to ensure they can be most effective in supporting you?

  Currently I can work around 3-3.5 hours daily, with the other commitment being college study. I have my final exams from May 27 to June 15th, so I think I'll be able to work a maximum of 2 hours daily during this period. However, my summer break starts after my exams end and goes on until August, so I can compensate for the exam period and contribute around 6-7 hours daily.

Appendix

My Github Profile - omkar334

My Linkedin Profile - Omkar Kabde

My Resume - Resume