

Motivational letter for 2020 ESiWACE Summer School - Lois Baker

As a PhD student on the Mathematics of Planet Earth CDT at Imperial College London, it is becoming increasingly clear to me that good computing skills will be essential if I am to continue my career in the field of physical oceanography.

I work with Dr Ali Mashayek on turbulent mixing processes in the deep ocean caused by the interaction of flow with topography. Whilst my background is in maths and is very theoretical, in the last year I have hugely increased my computing skills from near zero to a level where it is a part of my everyday work, and I hope to continue this throughout my PhD. In particular, I generally work with MATLAB, but am currently learning to work in Python as I know that this will be an essential skill in the future.

My need for HPC skills in my PhD work is twofold. I study small scale processes, and then try to make connections to global bulk energy and buoyancy budgets. For the former I set up and run simple simulations to study processes using the MIT general circulation model (MITGCM) and the Imperial HPC service, but currently lack the knowledge base to be able to customise these model runs and fully exploit post-processing tools. I would in particular like to learn more about data formats and data processing using Python.

For the latter, I have recently started using very large global climate model datasets - in my case this is to get predictions of velocities and stratification in the deep ocean to make predictions about the importance of different turbulent processes globally. In particular, I have started using Pangeo (pangeo.io) which is a community platform for big data geoscience, aimed at improving accessibility, reproducibility and scalability in geoscience. Pangeo is an incredible project, with some great tools for doing cloud computing in python notebooks hosted on Google Cloud. There are several huge open access datasets, including a 1/48 degree global ocean model, and Pangeo allows easy access using Xarray with Dask distributed clusters. It is a credit to this project that someone with relatively few computing skills can get started accessing this dataset, but in order to fully take advantage of what I firmly believe to be the future of climate science, I will need to hugely up-skill in big data analytics, storage, post-processing and visualisation.

I would propose a tentative project to use the Pangeo platform, since I think it is an idea that needs sharing widely. It would make a great group project, as there are several huge ocean/atmosphere datasets that are easily accessible, and everyone can freely access Jupyter notebooks with Dask Kubernetes compatibility. There are also many tutorial notebooks to help get started. There are plenty of interesting projects that could be proposed with this data - as an example, I'm interested in looking at energy flux in the high resolution ocean model between the mesoscale eddy field and bottom generated vortices in the ocean.

For these reasons, I think that this summer school would be hugely beneficial to me, my studies, and my group in Civil and Environmental Engineering. My fellow students, supervisor and I collaborate closely and all use oceanic data from observations and model runs. I have already been able to share the knowledge that I have gained in using Imperial HPC services to my group, through teaching my fellow students to access the services and writing a guide to running MITgcm using HPC. I share the opinion of my supervisor that it will be essential for our group to improve our research computing skills in the next few years to take advantage of the huge advances in open source big data. Through attending the summer school, I am hopeful that I can benefit both personally and by sharing the knowledge within my group and department.