Scientific computing and the Open Source software revolution: how Ocean Colour Science can benefit

Co-chairs: Joaquín E. Chaves (NASA GSFC/SSAI), Erdem M. Karaköylü (NASA GSFC/SAIC), & Joel P. Scott (NASA GSFC/SAIC)

Additional Discussion Leads: Dr. Myung-Sook Park (KIOST), Dr. Hayley Evers-King (Plymouth Marine Lab), & Bruce Monger (Cornell University)

Introduction

Until recently, ocean colour practitioners have principally relied on commercial off-the-shelf software (COTS) for data analysis (e.g., IDL, Matlab, etc). Use, support, and maintenance of COTS requires paid licenses, and often come with proprietary, black-box features. This framework hinders task-oriented modification and is an obstacle to transparency, code sharing, and scientific reproducibility. While COTS were instrumental in past progress toward better understanding the ocean and its processes, the restrictions associated with the use of COTS have become an obstacle to innovation and collaboration, hindering ocean colour science from truly realizing its full potential as a diverse global discipline of scientists, data users, and data producers.

During the past few years, there has been an explosive growth of information technology advances in computational power, data availability, and the open source software movement. These factors have resulted in the democratization of advanced computational tools and platforms for diverse commercial and scientific applications. There is now a rich ecosystem of accessible, open source software (OSS), that is freely available and modifiable, including programming languages, such as R, Python, Julia, and Octave. These tools are now easily accessible via the internet and their use is reinforced with online software and knowledge repositories such as GitHub, StackOverflow, Bitbucket, and others. OSS, combined with transparent scientific project management platforms like Slack and the Open Science Framework, have lowered the threshold for entry to ocean colour science, while expanding the user pool, increasing opportunity for collaboration, and promoting scientific innovation, transparency, and reproducibility. The rise of OSS enables new approaches for answering ocean colour research questions, conducting instrument calibration and algorithm validation, and streamlining data access, use, and availability.

Session Summary

Ocean colour science is a data-intensive discipline that requires advanced computational and analytical tools to fully realize its societal benefits. It is important to the ocean colour scientific community that the technologies being leveraged facilitate transparent, reproducible scientific results, while being accessible and understandable to allow for the training and mentoring of young scientists and new practitioners. As ocean colour science continues to expand globally, reliance on COTS has become a limiting factor. However, the growing adoption of OSS is encouraging collaboration that would otherwise have been a logistical impossibility. OSS and ‘open science’ principles are promoting diversity, inclusion, and accessibility to ocean colour science, driving international collaboration, and encouraging the key scientific principles of transparency and reproducibility. Examples of open source technologies and of how they are being leveraged to advance ocean colour research were highlighted in presentations and discussions centered around the following themes:

- Python is an emerging language with widespread adoption across all experience and skill-levels of the scientific community. Python offers a vast ecosystem of libraries that allow it to be a versatile choice for scientific computing with libraries for machine learning, modeling, data analysis/plotting, and web development.
- Jupyter notebook is a popular integrated development environment (IDE) that accommodates Python, Julia, and R open source languages and enables intuitive code testing and debugging features, as well as versatile tools for export and publication of code to encourage open, reproducible science.
- Open source Python modules are ideally suited for collaborative, scientific work, since they are 1) free from costly licensing, 2) easily installable in a ready-to-use state, and 3) well-supported with online resources, documentation, and tutorials. Some examples of Python modules used by the discussion leaders in this session, include:
    - NumPy for array computation
    - pandas for 2-D labelled data organization and manipulation
    - xarray for N-D labelled data manipulation with chunking and parallel processing
    - Matplotlib, Seaborn, and Cartopy for general, statistical, and geo-referenced graphing, respectively
    - PyMC3 for statistical modeling with explicitly debatable assumptions
- Anaconda is a popular Python package manager, used to handle module distribution and dependency resolution. Anaconda enables the creation of portable project-based environments, to track and export only the packages required for the project at hand (e.g., https://github.com/jpscot/IOCS_2019_Busan_OpenScience).

Recommendations

1. Develop and publish a community ‘open science’ statement to encourage making data and software open and discoverable. Responsible party: IOCCG

2. Encourage international adoption of ‘open science’ policies and open source technologies through existing training and education instances (e.g. - University of Maine Summer Ocean Optics class, EUMETSAT trainings, NASA SeaDAS trainings, Cornell Ocean Satellite class, etc). Responsible party: Agencies and the Community

3. Establish a code repository as a live IOCCG report, titled: Open Science Principles & Open Source Methods for Ocean Colour Science, to contain open source code and common ocean colour science workflows as a place-to-start for learning open source technologies. Responsible party: IOCCG to approve and host on GitHub/GitLab; Community/Agency members to contribute content and code examples