PACE Hackweek: A social coding event that keeps PACE with NASA's next great Earth science mission

A proposal for an OCB Training Activity

Jeremy Werdell, Anna Windle, Sean Bailey, Kelsey Bisson, Ian Carroll,
Sean Foley, Patrick Gray, Carina Poulin, Pengwang Zhai

Summary

We seek to host a social coding event (a.k.a. hackathon or codefest) that focuses on Earth science data streams from the upcoming NASA Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission. Our target audience is a diverse array of individuals from various backgrounds and career stages (students to professionals), with the dual goals of imprinting these researchers on our next generation mission while also improving global, public accessibility to its unprecedented data records. We will design the event with a combination of lectures and working (coding) activities that relate to OCB core objectives and host it at a minority serving institution. We anticipate providing access to a cloud-based compute platform, as well as tutorials on cloud computing, satellite and geospatial analysis in Python, usage of PACE data, and reproducibility. Participants will receive behind-the-scenes access to all aspects of the mission, including members of PACE with responsibilities for geophysical product implementation, instrument calibration, and data distribution. Qualified individuals from underrepresented groups in STEM will be prioritized to promote the diversity and inclusion aims of OCB and NASA. We are targeting August 2024, roughly 30 in-person students, accompanying virtual presentations, and a budget of $75,000. We will make all course material, including lecture and coding material, available for online dissemination.

Background and rationale

In 2015, NASA directed the PACE mission to Goddard Space Flight Center following recommendations from the 2010 NASA document *Responding to the Challenge of Climate and Environmental Change: NASA's plan for Climate-Centric Architecture for Earth Observations and Applications from Space*[1]. This direction ultimately realized the research communities' decade-plus push for a future Earth-observing satellite mission to meet growing needs for scientific discovery. A central objective of PACE is enabling new insights on the sensitivity of global aquatic ecology and biogeochemistry to environmental change. While heritage ocean color missions have provided desperately needed platforms for observing grossly under-sampled ocean ecosystems since 1997, the oceanographic community quickly recognized needs for enhanced satellite measurement capabilities to address the additional issues of changing phytoplankton distributions, ecosystem and habitat health, and carbon fluxes in the global oceans. A second driving objective of PACE is the retrieval of advanced atmospheric data products with the goals of reducing uncertainties in global climate models and improving our interdisciplinary understanding of the ocean-atmosphere system. Despite the substantial achievements of heritage satellite missions, better constraining aerosol and cloud properties and improving our understanding of effective radiative forcing requires

[1] https://science.nasa.gov/files/science-pink/s3fs-public/atoms/files/Climate_Architecture_Final.pdf

significant advances in measurement capabilities. Phrased simply, PACE represents NASA's next great investment in data records to enable continued and advanced insight into the changing Earth system.

Today, the PACE mission is part of the Program of Record in the National Academies of Science, Engineering, and Medicine 2018 Decadal Survey *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observations from Space*[2]. It supports a three-instrument payload: a global, hyperspectral (ultraviolet-to-shortwave infrared) scanning spectrometer (named OCI) focused on measurements of ocean color (aquatic biogeochemical products, such as phytoplankton community composition, and bio-optical properties) together with aerosols and clouds; and two small multi-angle polarimeters (named HARP2 and SPEXone) for measurements of detailed aerosol and cloud microphysical properties. OCI will enable never-before-seen observations of phytoplankton community dynamics, the polarimeters will provide unique observations of aerosol-cloud interactions (which are tied to the largest uncertainties in climate models), and the combination of the instruments offers a quantum leap forward in global Earth observations, yielding unprecedented insights to further humanity's understanding of our home planet and the climate stressors, ecosystem responses, and feedbacks already occurring.

In summary, the PACE mission emerged from a near two-decade-long science community effort aimed at **not only bringing satellite ocean color measurement capability *up to speed* with science capabilities and needs, but also providing data records of the Earth System that the next generation of scientist can *grow into***. And while it encompasses capabilities for both oceans and atmosphere, it also envisions synergistic opportunities between these science communities. The combination of hyperspectral, broad swath radiometric measurements from OCI (ranging from the UV to SWIR), hyperangular measurements from HARP2, and hyperspectral measurements from SPEXone substantially increase science value of the mission beyond what could be available from any one measurement or a single-sensor mission and offer never-before-seen looks at the Earth System. PACE is slated to launch in late January 2024 and will be transported to its launch site within the next month. Given this proximity to launch and beyond, **we believe the time is right to engage end-users, with the goal of empowering the community with the tools and know-how to access and interact with this next-gen data stream**.

Objectives

Our deliverable will be a one-week in-person hackathon that focuses on PACE's OCI, HARP2, and SPEXone data streams, with virtual and accompanying presentations to broaden the reach of the event. The objective for doing so is simple: a PACE-specific hackathon for researchers with various backgrounds and career stages (students to professionals) will help meet the dual goals of imprinting the community on this next generation mission, while also improving global, public accessibility to its novel data records. Building momentum behind engaged users of this data will be a multiplier on the investment in the mission. The event will encompass a balance of instructional and working time, with a time split of roughly 25% :75%, respectively. Lectures will focus on PACE science, data access, and software, as well as instruments' performance and how they relate to derived geophysical products, uncertainties, and

[2] https://www.nap.edu/catalog/24938/thriving-on-our-changing-planet-a-decadal-strategy-for-earth

usage (see Appendix A). Working time will involve rapid and collaborative efforts to create functional tools and improve existing software and open access code for data exploration and scientific analysis by the end of the event that addresses a series of PACE science-related projects, which can be pre- or attendee-defined (see Appendix B). We will have PACE Project members continuously involved to serve as lecturers and project/coding advisors and mentors, as well as to provide insight into details that that are often not readily available to the broader community (e.g., instrument and algorithm design choices that impact products and software version control rubrics) (see Appendix C).

Expected outputs

Our expected outputs are three-fold. First, we will conduct a one-week PACE data hackathon for ~30 in-person students (potentially more if the budget and logistics allow). Second, we will make all instructional materials and content generated for the hackathon publicly available online at the conclusion of the course (see Appendix A). In addition, lectures will be offered using a public online forum to allow non-participant remote attendance in real time. Third, and perhaps most important, we will unleash over two dozen empowered researchers and their newly developed code and analyses on the world. Participants will be educated about the benefits of cloud computing, but will also be capable of porting their projects from the cloud to on-premises servers or local machines. The latter will not only assist the participants with future career endeavors, but also support interdisciplinary research and increase the accessibility and use of data from the unique and advanced PACE observatory, which should ultimately improve our scientific understanding of our home planet.

Support for OCB priorities

Our proposed training activity directly address two core objectives of the OCB program:

- *"... explore the ocean's role in the global carbon cycle and the response of marine ecosystems to environmental changes of the past (paleo), present, and future (prediction)*.": The primary promise of PACE is an advanced scanning spectrometer to substantially improve capabilities for the remote retrieval of phytoplankton community composition and carbon stocks. Its two multiangle polarimeters will offer additional potential for retrieving microphysical hydrosol properties (e.g., aquatic particle size distributions and refractive indices) as well as offer a remarkable opportunity to better understand the coupled atmosphere-ocean system, which significantly influences the global carbon cycle. In summary, the outcomes of our proposed hackathon are in direct alignment with facilitating this core objective of OCB.

- *"Develop education and outreach activities and products with the goal of promoting ocean carbon science to broader audiences."*: The core objective of the proposed hackathon is to advance accessibility and operability of passive remote sensing data records of the Earth System. In addition, our proposed event will offer opportunities for attendees to meet, be mentored by, and forge relationships with PACE experts, as well as for attendees to network amongst

themselves and develop novel research concepts. In summary, the *raison d'être* for our proposed activity is in direct alignment with this core objective of OCB.

Participation

Participants: We envision accepting applications for roughly 30 in-person enrollments. As a course tailored to include students as well as professionals, the class will be held in the summer and we expect to start accepting applications in early 2024. We expect the application package to include: (1) CV, including details on programming experience with expectations of a baseline competency; (2) personal statement, including commentary on potential projects to be conducted during the event; and, optionally, (3) a general reference letter if possible and as appropriate. We hope to include both U.S. and international students. To promote the diversity and inclusion aims of OCB and NASA, we will prioritize qualified individuals from underrepresented groups in STEM for acceptance into the course and utilize NASA's MUREP[3] (and similar) mailing lists for recruitment. To reduce a barrier to travel-for-work that is often unequally borne by underrepresented groups in STEM, we will work with OCB to explore options for reimbursing "at-home" dependent care expenses or provide hotel accommodation for participants who need to travel with dependents. These benefits will be advertised in the course application, and acceptance letters will advise participants to describe their dependent care needs.

Additional community involvement: We envision providing all lectures using a public online forum that allows non-in-person parties to watch and engage with moderated Q&A in real time (e.g., WebEx, Swapcard, YouTube). We will also pursue a venue-wide seminar on PACE to students/faculty at the host institution during the hackathon week and a post-class virtual seminar to the interested public on the outcomes of the class and how all can get involved and benefit from the class deliverables.

Instructors/mentors: We envision a dynamic array of participation in this category, inclusive of the mission Project Scientists and their team leads, NASA programmers and data archival specialists, and community experts in cloud computing, to name only a few. A list of potential instructors is provided in Appendix C. This list of mentors represents a variety of career stages that will encourage a flat hierarchy in terms of culture. We anticipate participant/mentor ratios of 3:1 to 5:1, pending mentor availability. Several mentors have been involved with similar hackweek workshops (Appendix C), providing valuable information on working through potential problems.

Logistics and budget

Location: While lectures can be made virtual, working time will be in-person only. We envision the in-person component will be conducted as follows: We will host this class at the University of Maryland Baltimore County (UMBC), a minority serving institution in Baltimore, Maryland, and the venue of the successful OCB-sponsored 2022 PACE class. UMBC has the facilities and infrastructure to host the training event's classes and laboratories, lodge and house participants and provide meals. As a bonus,

[3] Minority University Research and Education Project; https://www.nasa.gov/stem/murep/home/index.html

UMBC is also contributing the HARP2 multi-angle polarimeter that will fly on the PACE observatory, which will provide additional insider knowledge on its development and science capabilities. If UMBC as hosts cannot be realized, a potential backup venue will be the Welcome Center at GSFC in Greenbelt, Maryland, which includes classrooms and offers reasonable access to local hotels and restaurants and the DC Metro train and bus system. As for the previous class, evening activities, such as "fireside chats", local experiences or meals (crabs), and relevant field trips to NASA or elsewhere will be planned.

Computing: The Ocean Biology Distributed Active Archive Center (OB.DAAC) will distribute PACE data through NASA's Earthdata Cloud on AWS S3, the first instance of OB.DAAC data being distributed through a commercial cloud provider. Participants will learn about the new computing approaches made possible with Earthdata Cloud by doing all hackweek activities in an AWS EC2 instance co-located with PACE data. The entry point will be a JupyterHub site, for which we expect support from NASA. Coding will be encouraged to be primarily in Python and only rely on free and open-source software. While everything taught may work best in the cloud, it will also all work while not on the cloud. Participants will need to provide a laptop with a modern web-browser, as all coding and analysis will be done in a browser-based JupyterLab session running on this server. Following the end of the hackweek, attendees will be able to run an identical development environment either locally on their machines or on their own cloud server by pulling the image from DockerHub. Lectures will be streamed as possible and recorded. Tentative lecture concepts are provided in Appendix B. We will use GitHub as our primary content delivery platform and publish links to online notebooks that participants create.

Budget: We request $75,000 to host this event. Scoping costs are based on the UMBC rates proposed for the OCB-supported PACE class in 2022 and are only used here for preliminary budgeting. The following considers a single one-week event that includes 30 students and 12 mentors:

Host facility cost	$15,000
Mentor food & catering (12 mentors x 5 days x $60/day)	$3,600
Van rentals for field trips	$2,500
Travel & hotel for 4 non-local instructors	$6,000
Participant accommodations (30 students x ~$1,490)	$44,700
Draft total for 30 students and 12 mentors	**$71,800**

The per participant cost of ~$1,465 breaks down as follows: $390 for 6 days of single room dorm housing, $300 for 5 days of food and catering, $600 for airfare, and $200 for ground transportation. Note that any remaining funds from the requested $75,000K can be used to enlist additional participant, support participant incidentals or meals during travel days, provide additional local experiences or field trips, recruit additional non-local mentors, and otherwise be used to maximize the participant and class experience. Funds are not requested to support local mentors beyond the costs of their meals.

Timing: We plan to tentatively host this class in the 5-16 August 2024 timeframe. Given student schedules, the class could likely proceed one week later, but would otherwise need to slip to a winter/semester break timeframe, e.g., January 2025, to ensure some desired student attendance.

BONUS MATERIAL

Appendix A. Potential lectures.

The following are general topics we feel would be of value to cover:

- Overview of PACE
- Ocean color, polarimetry, and remote sensing 101
- PACE data architecture - netCDF, metadata, data processing levels, standard products
- PACE data accessibility and access methods
- Git and github
- Xarray in 45 minutes
- Geospatial data analysis and visualization tools
- Software environment management (Python packages and shared libraries)
- Uncertainties for performance assessments
- Machine learning
- Cloud computing
- Beyond the hackathon – how to continue working with PACE data

The instructional portions of the week will be broken up with two lectures each day, with one focusing on science topics and one focusing on computing topics. The lectures will be high-paced, condensing a lot of information that individuals can subsequently unpack with their project teams and mentors during work periods. The instructional sessions may be held in a separate facility, so that participants who prefer to continue project work can do so without interruption.

The following provides an example of instructional lecture workflow. The exact content, cadence, and schedule will be defined later – this table is provided only to be illustrative of the workflow.

Monday	PACE and PACE Data - overview of PACE instruments - overview of ground segment - overview of data products
	Cloud Orientation and `earthaccess` - When & why of commercial cloud data storage and computation - Orientation to the PACE Hackathon's JupyterHub - Getting data with the `earthaccess` package
Tuesday	Bio-optical inversions of hyperspectral data
	Manipulation and Visualization of Multidimensional Array Data - XArray in under 45 minutes - Holoviews or placing arrays in visualizable containers
Wednesday	Polarimetry Lecture
	Collaboration via Repositories - git repositories for code - options for "repositories" for data - collaboration through github

Thursday	A Wunderkammer of PACE Algorithms - get to know them all (a little bit)
	Performant Computing - dask and other paralellization tools - how to troubleshoot, debug - how to evaluate performance, profile
Friday	Uncertainties/Calibration/Validation Lecture
	Re-usable Projects - re-usable means portable, shareable, and reproducible - tools for encapsulating the compute environment (venv, docker, conda) - options for accessing the commercial cloud, or not

Appendix B. Potential project concepts

Projects will ultimately be crafted based on participant input.

Projects could include the following concepts tailored to emerging questions or participant interest:

- Comparative trend analyses on multiple scales or platforms
- PCA analyses using hyperspectral vs. multispectral information
- PCA analyses using radiometry, polarimetry, and/or external environmental information
- Spectral derivative analyses using hyperspectral data
- Correlative analyses with environmental variables from in situ assets or assimilation models
- Closure analyses with in situ assets or Earth system/assimilation models
- Data merger or intercomparisons across multiple instrument or satellite platforms
- Biogeochemical clustering of data into seascapes, provinces, etc.
- Algorithm and retrieval methods intercomparisons
- Rare colors/spectra identifications; Evaluation of "150 shades of green"
- Polarimetric retrievals for aquatic applications
- Uncertainties estimations and cross-platform uncertainties validation

Appendix C. Potential list of instructors / lecturers / mentors

Lecturers will ultimately be defined as the agenda is crafted.

The following provides a list of committed participants with responsibility for event planning:

- Sean Bailey (GSFC; OB.DAAC Manager)
- Kelsey Bisson (NASA HQ; Oregon State University) – *Participated in the 2020 ICESat-2 hackweek, served as planning committee member for 2022 ICESat-2 hackweek*
- Ian Carroll (GSFC; PACE scientist) – *Lead instructor for the SESYNC Summer Institute on Cyberinfrastructure 2016–2019*

- Sean Foley (GSFC; PACE scientist)
- Patrick Gray (University of Maine) - *Participated in 2020 OceanHackWeek and 4 other hackathons, designed and taught two undergrad courses on remote sensing and machine learning on JupyterHub*
- Carina Poulin (GSFC; PACE outreach and scientist)
- Jeremy Werdell (GSFC; PACE Project Scientist)
- Anna Windle (GSFC; PACE scientist)
- Pengwang Zhai (UMBC; Associate Professor)

The following provides a list of potential (identified but to-be-contacted) supporting instructors/mentors:

- Riley Blocker (GSFC; PACE scientist)
- Guillaume Bourdin (University of Maine)
- Brian Cairns (GISS; PACE Deputy Project Scientist – Atmospheres)
- Ivona Cetinić (GSFC; PACE Project Science Lead for Biogeochemistry)
- Alison Chase (University of Washington)
- Zach Fair (GSFC)
- Nils Haentjens (University of Maine)
- Kirk Knobelspiesse (GSFC; PACE Project Science Lead for Polarimetry)
- Sasha Kramer (MBARI)
- Amir Ibrahim (GSFC; PACE Project Science Lead for Atmospheric Correction)
- Antonio Mannino (GSFC; PACE Deputy Project Scientist – Oceans)
- Romina Piunno (University of Toronto)
- Andrew Sayer (GSFC; PACE Project Science Lead for OCI Atmospheres)
- Jessica Scheick (University of New Hampshire)
- Alicia Scott (GSFC; OB.DAAC Deputy Manager)
- Emerson Sirk (GSFC; PACE scientist)