The PyHC Open Science Experiment: A PyHC session led by Rebecca Ringuette

Rebecca Ringuette¹

¹Affiliation not available

May 25, 2023
The PyHC Open Science Experiment

Rebecca Ringuette and many others

NASA GSFC Center for HelioAnalytics

Presented at the 2023 PyHC Spring Meeting at LASP/CU Boulder in Boulder, CO.
Session Outline

• Invited presentation on the **Open Science Framework** (10-15 min)
  Presented by Gretchen Geugeun

• **Project Introduction** (15-20 min)
  • The PyHC Open Science Experiment
  • Project Tour

• **Open Science and Heliophysics Infrastructure** (5-10 min)

• **Discussion** (45 min): What PyHC software changes are needed to better support this project and, more generally, open science? What funding is needed to complete these tasks?
The Open Science Framework
Presented by Gretchen Geugeun
Link to slides: https://docs.google.com/presentation/d/1vtSmbsDweTLm8Sgw6GfwNu9aDQM3iMnPoQTlKje_M/edit?usp=sharing
Project Introduction: The PyHC Open Science Experiment

• The PyHC executable paper demonstrated:
  • How to collaborate between software developers/engineers and scientists,
  • How to use multiple PyHC packages to perform a science analysis,
  • How to produce an executable paper in Heliophysics, and
  • How such a collaboration supports open science.

• The goals for this work are to:
  • Apply the workflow developed to a full-scale science problem, specifically expanding the 2015 challenge with new data from MMS (https://ccmc.gsfc.nasa.gov/challenges/gem-magnetopause/),
  • Demonstrate how to perform open science in Heliophysics, and
  • Improve and develop modern infrastructure to streamline collaboration and contributions.
Project Introduction: The PyHC Open Science Experiment

GEM Science Plan

• Expand to include multiple time ranges of MMS data where magnetopause crossings occurred (retrieved with pySPEDAS),

• Generate the predictions using the empirical Shue model (using SpacePy),

• Generate flythrough results for each contributed physics-based model output stored in s3 (via Kamodo),

• Encourage the community to provide metrics calculation scripts using the flythrough results (built on PlasmaPy), and

• Provide a platform where all contributors can search and reuse all components (on HelioCloud).

• Multiple members of the community are expected to lead portions of the project and produce multiple papers, including a summary paper (coordinated on the Open Science Framework).

NEED SOME EXAMPLES!
Project Introduction:
The PyHC Open Science Experiment

Open Science goal

• Perform the work in the open from the beginning,

• Demonstrate how to perform open science to the Heliophysics community and various agencies and nations,

• Develop any lacking infrastructure along the way (as reasonably possible),

• Create examples of rubrics for recognition/coauthorship and contribution/participation rules for open science, and

• Publish a paper describing the challenges discovered, lessons learned, advancements achieved, and how this work can be expanded upon.
Project Tour: OSF Project Page

Comparing Magnetopause Crossings using Open Science

Contributors: Rebecca Ringuette
Date created: 2023-05-08 09:29 AM | Last Updated: 2023-05-11 06:39 PM
Identifier: DOI 10.17605/OSF.IO/V4DRT
Category: Project

Description: Poisson et al. (2022, https://doi.org/10.3388/psp.2022-977781) successfully demonstrated how to build an executable paper by collaborating between research software engineers, software developers, and scientists in heliophysics for the first time. In their work, a single MMS observation of a magnetopause crossing was compared with the Shue model and a simulated result from the OpenGCM model, all using Python packages from the PyHC community (https://helipython.org/). The result is an executable paper hosted on DeepNote for the community to build directly upon, which is linked in the published paper. That work was limited by a data volume of 2 GB, which was not enough to compare the observed results with modeled data of sufficient quality. Also, only one model output from one model was supported by the versions of software packages included in that work.

This work aims to build that executable paper into an open validation platform as an example of how the application of open science principles can accelerate science. The analysis platform will be hosted on HelioCloud, an online analysis platform under development in Heliophysics, where the required software and developed infrastructure will be made public. The project page on OSF will serve as the main resource with non-technical resources, such as contribution rules and recognition rubrics, and links to HelioCloud, the associated GitHub page, and other resources. When the project is made public, the community will be able to openly contribute resources (e.g., model outputs, scripts, and software tools) to the platform and use any contributed object openly. The platform is expected to be made public sometime in late 2023.

License: Apache License 2.0

Project web page: osf.io/v4drt/ DOI: 10.17605/OSF.IO/V4DRT

Please make OSF/ORCiD accounts so I can add contributors!
Project Tour: HDRL’s HelioCloud

- Cloud computing environment
- Executable and shareable notebooks
- Large file storage supported via public s3 buckets
- Initial compute and storage costs funded by HDRL

https://daskhub.hsdcloud.org

...more tomorrow in S. Antunes’ presentation.
Project Tour: GitHub page

• Link to project webpage added to readme file.
• Scripts and notebooks stored in ‘DataWorkflows’ folder.
• Software environment information in top directory.

https://zenodo.org/badge/latestdoi/631044088
Project Tour: Linking It All Together

People on the OSF pages will browse the data in s3 buckets and the files on HelioCloud through an intuitive interface.

Contributors will perform all data analysis on HDRL’s HelioCloud.

People on the OSF pages will also see software, discussions, documentation and contributors embedded from GitHub.

Contributors will easily push/pull software between GitHub and HelioCloud.

Main Project Page

Main GitHub Repo

Forked GitHub Repos

OSF

Project Component Pages

HelioCloud

git
Project Tour: Current Status

- Project posted on OSF with a DOI
- Dependency conflicts resolved on HelioCloud
- **Software runs on data in s3!** (except for SWMF GM outputs)
- Workflows being planned and developed
Project Tour: Path Forward

- **Work out kinks** in running PyHC software on data in s3 buckets.
- **Add Contributors!** *(Make an account on OSF/ORCiD so I can add you!)*
- **Link** HelioCloud, OSF and GitHub together (easier said than done!).
- **Streamline** workflows for contributors *(NEED TESTERS!)*.
- **Draft** contribution/participation rules based on JWST example.
- **Finalize** contribution/participation rules at Fall PyHC meeting.
- **Present** at AGU 2023 a (hopefully) ready environment.

**Any burning questions?**
Open Science and Heliophysics Infrastructure

- **Open Science** is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity. ([https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/](https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/))

- **Why open science?**
  - Accelerates scientific discovery.
  - Greater collaboration and efficiency.
  - Enhanced transparency and reproducibility ([NASEM, 2018, p. 3](https://zenodo.org/record/6565080#.ZFPvCnbMKUk)).
  - Mandated by the U.S. White House and NASA.

- Image Credit: NASA TOPS ([https://zenodo.org/record/6565080#.ZFPvCnbMKUk](https://zenodo.org/record/6565080#.ZFPvCnbMKUk))
Open Science and Heliophysics Infrastructure

• FAIR components (Findable, Accessible, Interoperable, Reusable):
  • Making good progress: publications, observed data, metadata,
  • Needs focused development: modeled data, software,
  • Exploration required: model codes, software environments,
  • The great unknown: people, relationships, collaborations, ...

• Reproducible results
  Executable papers? Analysis environments? How long to maintain and what depth of reproducibility?

• Open processes
  How to perform science in the open from the beginning?

• Inclusive collaborations
  How to make collaborations open?

FAIR data and open-source software are NOT enough!
It is okay to start there, but we must look beyond for guidance on infrastructure design.
Observational Data

• A growing number of datasets...
  • Are **searchable** through a modern interface (using SPASE),
  • Have **citable** DOIs independent of publications,
  • Are **downloadable** both through web pages and APIs,
  • Are **browsable** via quick-look plots, and
  • Are available on the **cloud**.

**How can PyHC better advertise data access and analysis support in PyHC packages?**
Modeled Data

- Infrastructure supporting modeled data is **far less developed**.
  - No modern search interfaces,
  - DOIs are not assigned,
  - Few modeled datasets are **downloadable** through a website,
  - Only reduced versions are **available** through an API, and
  - Some quick-look **plotting** capabilities are available, but are not easily accessible.

*How can PyHC help with these issues?*
Software

• Sustained push is underway to **open-source all software** (including modeling code) generated with taxpayer dollars.

• Open-sourcing software is **NOT enough**.
  • Dependency conflicts (!),
  • Conda/pip installability on multiple operating systems (e.g. Mac, Windows, Linux),
  • Lacking documentation,
  • Need examples and tutorials,
  • Capability to run on the cloud,
  • Maintenance for long-term reusability,
  • Support staff for questions/problems, and
  • Containerization for software environments?

*PyHC should take the lead here. What paths forward have low-hanging fruit?*
Where is this going?

• Build a **distributed data infrastructure** system:
  • Observational and modeled data hosted and served by multiple institutions,
  • Containerized model codes available on the cloud from multiple institutions,
  • All searchable from a united modern interface through **connected metadata**,
  • All accessible using multiple methods (e.g. file links, APIs, quick-look plots).

• Build a **collaborative analysis infrastructure** system:
  • Analysis environments with **software already installed** (and referenceable),
  • Reusable executable analysis tutorials for how to use the data,
  • Searchable through **connected metadata**,
  • Accessible through the cloud (e.g. downloadable containers or cloud platforms).

*How can PyHC prepare for, collaborate with, and enable these infrastructures?*
Discussion Time!
Discussion: PyHC support of Open Science
Discussion: PyHC support of Open Science

Scan the QR code or use the link to contribute to the discussion!

https://tinyurl.com/5n6tsxt6
What PyHC software changes are needed to better support the open science experiment? Which of these need funding?

- How to run SWAP 3D data stored in s3 buckets?
- Need a metrics script that uses PlasmaPy
  - mean error, RMS, root square mean, absolute error

PyHC packages should test against the main branch or release candidates of their dependencies.

PyHC packages should implement best practices.

A common release schedule so that there is always a compatible / functioning, relatively recent and complete set of PyHC tools.

We need funding for long-term maintenance of code!

acknowledgment of the code in papers

understand/development of the metric of the code “usefulness”

DOI’s for software - should PyHC offer DOI minting or should a DOI be required for software to be listed by PyHC?

PyHC office hours?

Standardized packaging - packages install scripts dir structure

How can PyHC better advertise data access and analysis support in PyHC packages?

- Generate a SPASE metadata record for software to better enable discoverability at MDLP?
- Create more links on the archive page to relevant subtask?
- Create a SPASE metadata record for software to better enable discoverability at MDLP?
- Create SPASE metadata record for software to better enable discoverability at MDLP
- Improve keyword usage/search on PyHC Projects page
- Out-of-the-box: A custom GitHub-powered cheat sheet on PyHC that can explain what's available in conversation
- Community cheat sheets with examples of how to use a package - someone or a team to use a new package, the package code itself (with some curations) could be useful examples for the next release?

Webinars (similar to pySPEDAS)

Advertisement at ALL conferences

Do we have a standard set of tutorials that anyone going to a meeting can take so that we have lots of representatives who do the training?

PyHC should have an array of tutors - offer specific PyHC grants for people to do promotion - or this could be an add-on to another grant - a little extra money to attend one or two extra meetings and give tutorials

Should PyHC hire a community manager like Astropy?

Hosting work parties to improve project documentation

To include these in the PyHC gallery?

booth with tutorials at meetings

Community cheat sheets with examples of how to use a package - some or a team to use a new package, the package code itself (with some curations) could be useful examples for the next release.
Useful Links

• OSF project page: https://osf.io/v4drt/
• GitHub project page: https://github.com/rebeccaringuette/MagnetopauseExecutablePaper or https://zenodo.org/badge/latestdoi/631044088
• HelioCloud: https://daskhub.hsdcloud.org
• CCMC 2015 Magnetopause challenge: https://ccmc.gsfc.nasa.gov/challenges/gem-magnetopause/
• Polson et al. 2022 journal article: https://doi.org/10.3389/fspas.2022.977781
• Polson et al. 2022 on Deep Note: https://deepnote.com/workspace/shawn-polson-c095a0fb-f02d-416d-9c94-c4a9c4e8e54d/project/PyHC-Paper-101b9646-3fd0-4978-a48e-a4f3e708a0ac/notebook/Making_an_Executable_Paper_with_the_Python_in_Heliophysics_Community_to_Foster_Open_Science_and_Improve_Reproducibility-c3a572e5ce24ce15426b001696d52251
• This presentation’s link on PyHC’s google drive: https://docs.google.com/presentation/d/1c2bP0zOoDjYWMCPZ2Zxm9U3x80NPf_SC/edit?usp=share_link&ouid=11819833928784120748&rtpof=true&sd=true
• Miro board link with preliminary project workflows (view): https://miro.com/welcomeonboard/Q05NeGiz2MTaVGxtexkjjZXkzVlsdzFud3R3SF4RWJ4RG0NXBMazZheThhd1d4aUNLc1VD0F1W/HhXa01ZWXwzMDc0NDU3MzU3OTk0ODEcyOTEyfDi=？share_link_id=953968385486
• Miro board link with PyHC 2023 spring session discussion (edit): https://miro.com/welcomeonboard/M21Rd22NSXxPcXpPRXZJbDHFHczcwRHNgzyUzSk9XelBzeVdkt72p3NTHhQlVBKg4VJeaaXc5MWNGT3dhUkxR3wzMdc0NDU3MzU3OTk0ODEcyOTEyfDi=？share_link_id=522570131865
Project Tour: Preliminary Project Workflows

https://miro.com/app/board/uXjVMMaO61I=/?share_link_id=703942255605
Project Tour: Preliminary Project Workflows

https://miro.com/app/board/uXjVMMaO61I=/?share_link_id=703942255605