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

Rebecca Ringuette[1]

[1]Affiliation not available

August 3, 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/1vtSmbsDweTLmS8wGgfwNu9aDQMZ3iMnPQeQTlKJje_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***).

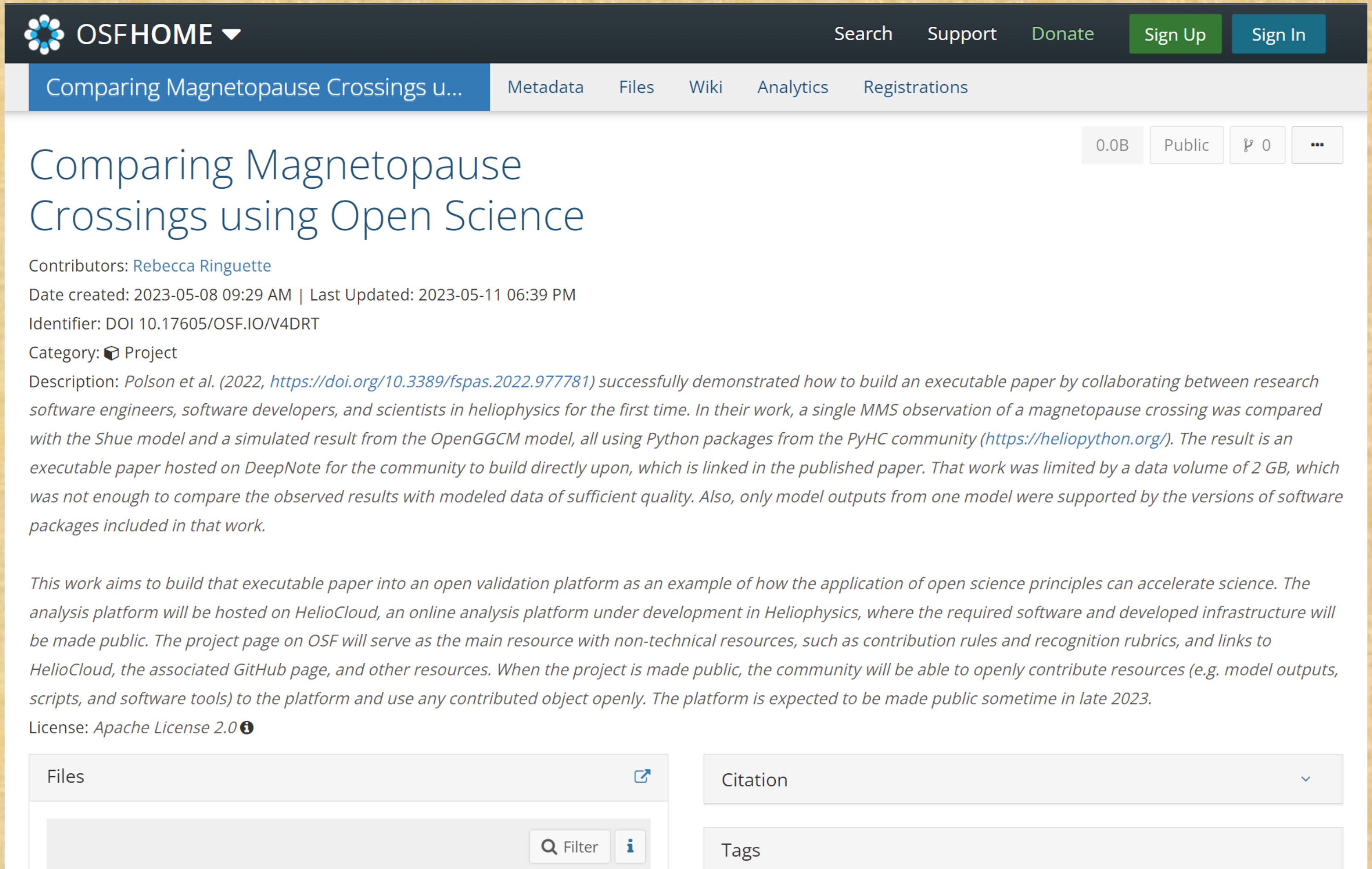
NEED SOME EXAMPLES!

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: *Polson et al. (2022, https://doi.org/10.3389/fspas.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 OpenGGCM model, all using Python packages from the PyHC community (https://heliopython.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 model outputs from one model were 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*

Please make OSF/ORCiD accounts so I can add contributors!

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

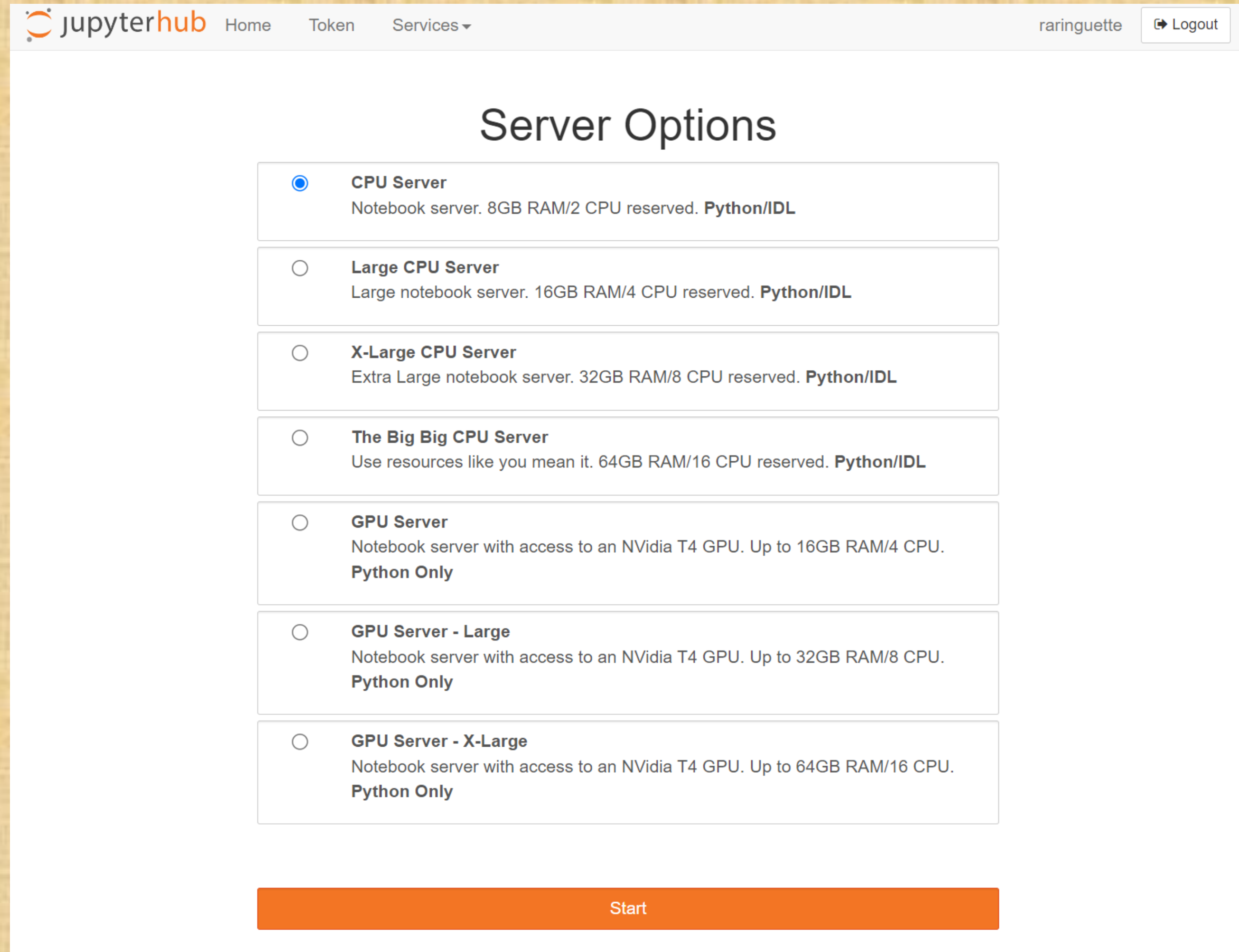
Project Tour: HDRL’s HelioCloud

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.



Server Options

● **CPU Server**
Notebook server. 8GB RAM/2 CPU reserved. **Python/IDL**

○ **Large CPU Server**
Large notebook server. 16GB RAM/4 CPU reserved. **Python/IDL**

○ **X-Large CPU Server**
Extra Large notebook server. 32GB RAM/8 CPU reserved. **Python/IDL**

○ **The Big Big CPU Server**
Use resources like you mean it. 64GB RAM/16 CPU reserved. **Python/IDL**

○ **GPU Server**
Notebook server with access to an NVidia T4 GPU. Up to 16GB RAM/4 CPU. **Python Only**

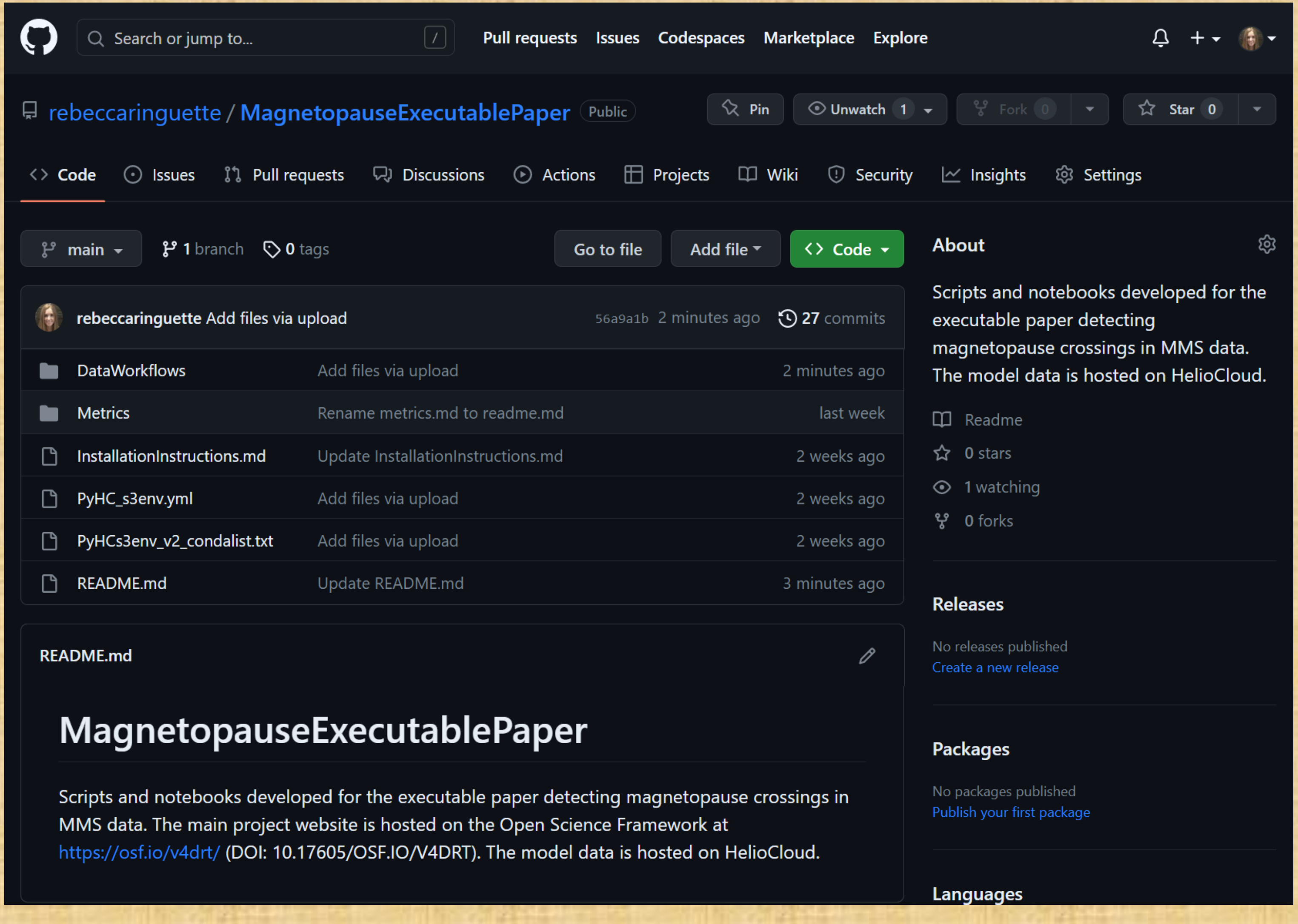
○ **GPU Server - Large**
Notebook server with access to an NVidia T4 GPU. Up to 32GB RAM/8 CPU. **Python Only**

○ **GPU Server - X-Large**
Notebook server with access to an NVidia T4 GPU. Up to 64GB RAM/16 CPU. **Python Only**

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



rebeccaringuette Add files via upload 56a9a1b 2 minutes ago 27 commits

DataWorkflows	Add files via upload	2 minutes ago
Metrics	Rename metrics.md to readme.md	last week
InstallationInstructions.md	Update InstallationInstructions.md	2 weeks ago
PyHC_s3env.yml	Add files via upload	2 weeks ago
PyHCs3env_v2_condalist.txt	Add files via upload	2 weeks ago
README.md	Update README.md	3 minutes ago

README.md

MagnetopauseExecutablePaper

Scripts and notebooks developed for the executable paper detecting magnetopause crossings in MMS data. The main project website is hosted on the Open Science Framework at https://osf.io/v4drt/ (DOI: 10.17605/OSF.IO/V4DRT). The model data is hosted on HelioCloud.

About

Scripts and notebooks developed for the executable paper detecting magnetopause crossings in MMS data. The model data is hosted on HelioCloud.

Readme
0 stars
1 watching
0 forks

Releases

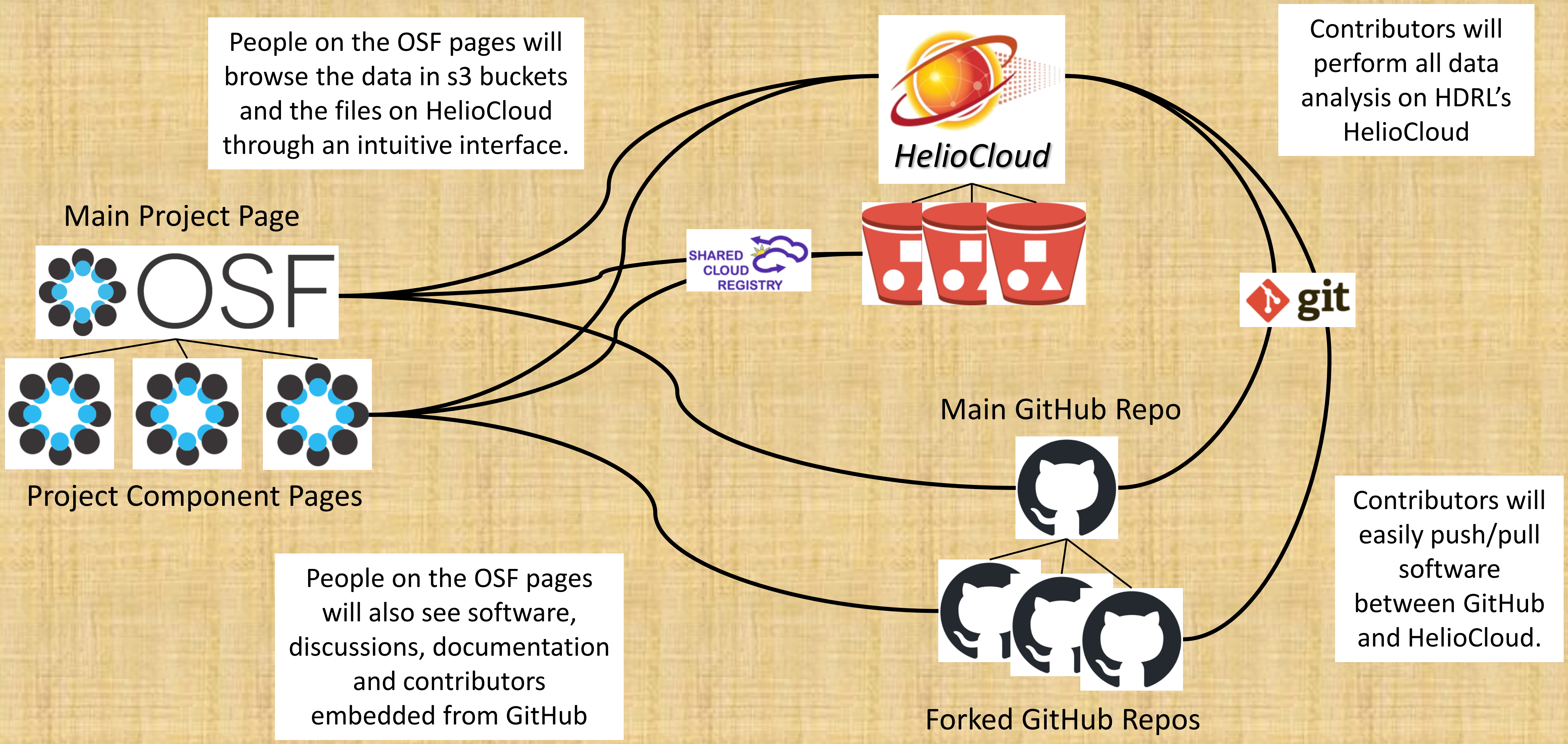
No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

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.

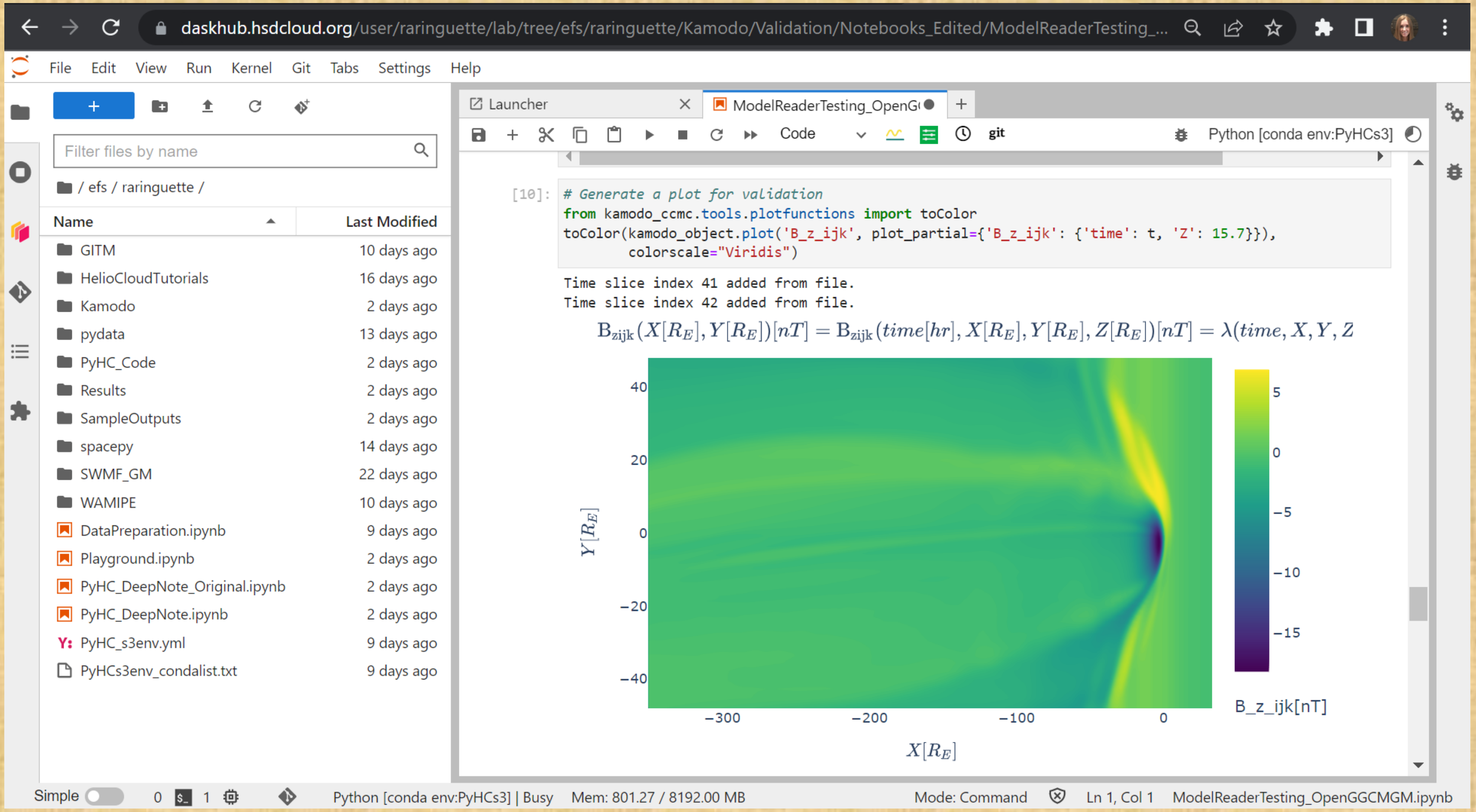
Contributors will perform all data analysis on HDRL’s HelioCloud

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

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

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



daskhub.hsdcloud.org/user/raringuette/lab/tree/efs/raringuette/Kamodo/Validation/Notebooks_Edited/ModelReaderTesting_...
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Filter files by name
/ efs / raringuette /
Launcher
ModelReaderTesting_OpenG
Code
git
Python [conda env:PyHCs3]

Name	Last Modified
GITM	10 days ago
HelioCloudTutorials	16 days ago
Kamodo	2 days ago
pydata	13 days ago
PyHC_Code	2 days ago
Results	2 days ago
SampleOutputs	2 days ago
spacepy	14 days ago
SWMF_GM	22 days ago
WAMIPE	10 days ago
DataPreparation.ipynb	9 days ago
Playground.ipynb	2 days ago
PyHC_DeepNote_Original.ipynb	2 days ago
PyHC_DeepNote.ipynb	2 days ago
PyHC_s3env.yml	9 days ago
PyHCs3env_condalist.txt	9 days ago

[10]:

```python
# Generate a plot for validation
from kamodo_ccmc.tools.plotfunctions import toColor
toColor(kamodo_object.plot('B_z_ijk', plot_partial={'B_z_ijk': {'time': t, 'Z': 15.7}}),
        colorscale="Viridis")
```

```
Time slice index 41 added from file.
Time slice index 42 added from file.
```

$$\mathrm{B_{zijk}}(X[R_E], Y[R_E])[nT] = \mathrm{B_{zijk}}(time[hr], X[R_E], Y[R_E], Z[R_E])[nT] = \lambda(time, X, Y, Z$$

Simple 0 1 Python [conda env:PyHCs3] | Busy Mem: 801.27 / 8192.00 MB Mode: Command Ln 1, Col 1 ModelReaderTesting_OpenGGCMGM.ipynb

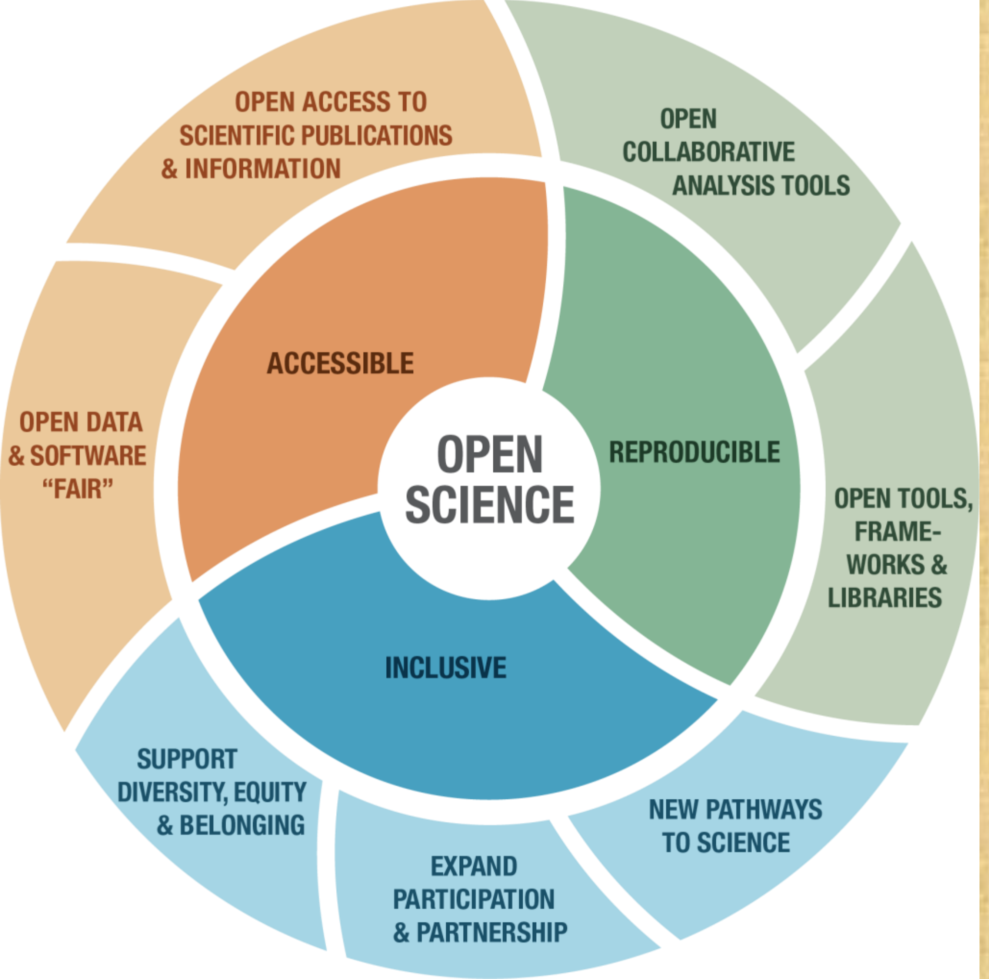
Project Tour: Path Forward

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


OPEN ACCESS TO SCIENTIFIC PUBLICATIONS & INFORMATION
OPEN COLLABORATIVE ANALYSIS TOOLS
OPEN DATA & SOFTWARE "FAIR"
ACCESSIBLE
OPEN SCIENCE
REPRODUCIBLE
OPEN TOOLS, FRAME-WORKS & LIBRARIES
INCLUSIVE
SUPPORT DIVERSITY, EQUITY & BELONGING
NEW PATHWAYS TO SCIENCE
EXPAND PARTICIPATION & PARTNERSHIP

Open Science and Heliophysics Infrastructure

- **FAIR** components (***F**indable, **A**ccessible, **I**nteroperable, **R**eusable*):
  - *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?

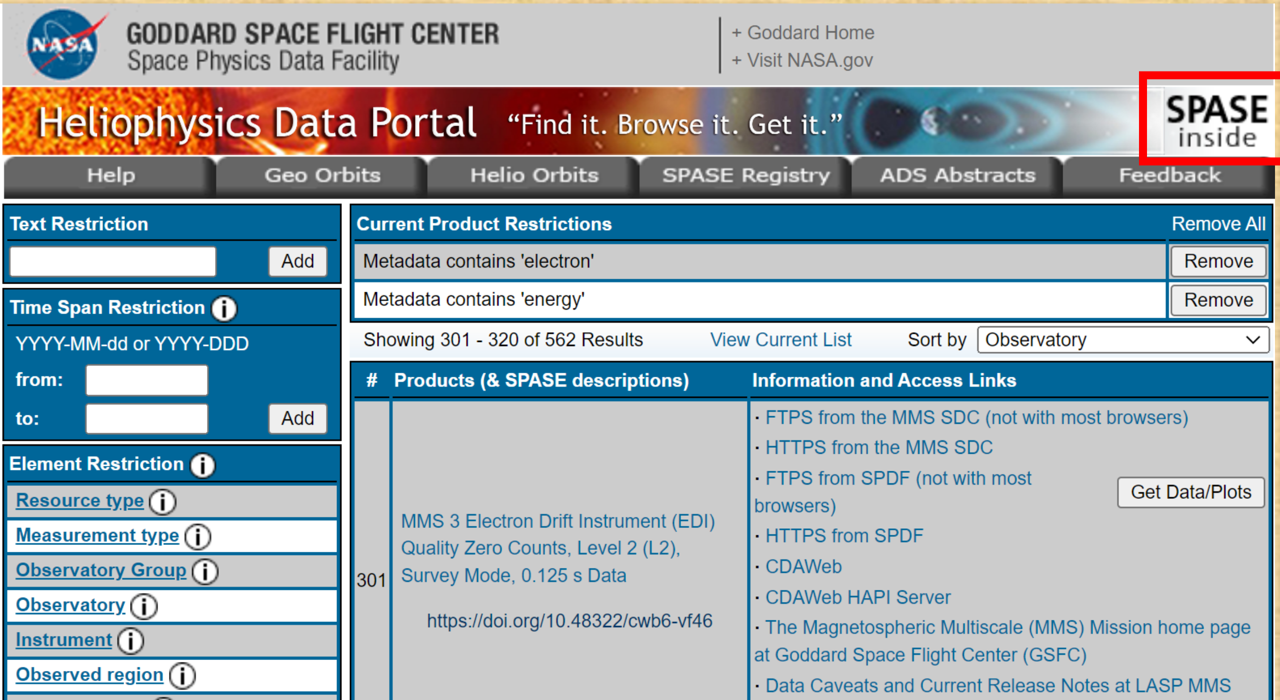
FAIR data and open-source software are NOT enough!

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?



GODDARD SPACE FLIGHT CENTER
Space Physics Data Facility

+ Goddard Home
+ Visit NASA.gov

Heliophysics Data Portal "Find it. Browse it. Get it."

SPASE inside

Text Restriction [ ] Add

Time Span Restriction ⓘ
YYYY-MM-dd or YYYY-DDD
from: [ ]
to: [ ] Add

Element Restriction ⓘ
Resource type ⓘ
Measurement type ⓘ
Observatory Group ⓘ
Observatory ⓘ
Instrument ⓘ
Observed region ⓘ
Spec
Cade
Repo
Acce
Form

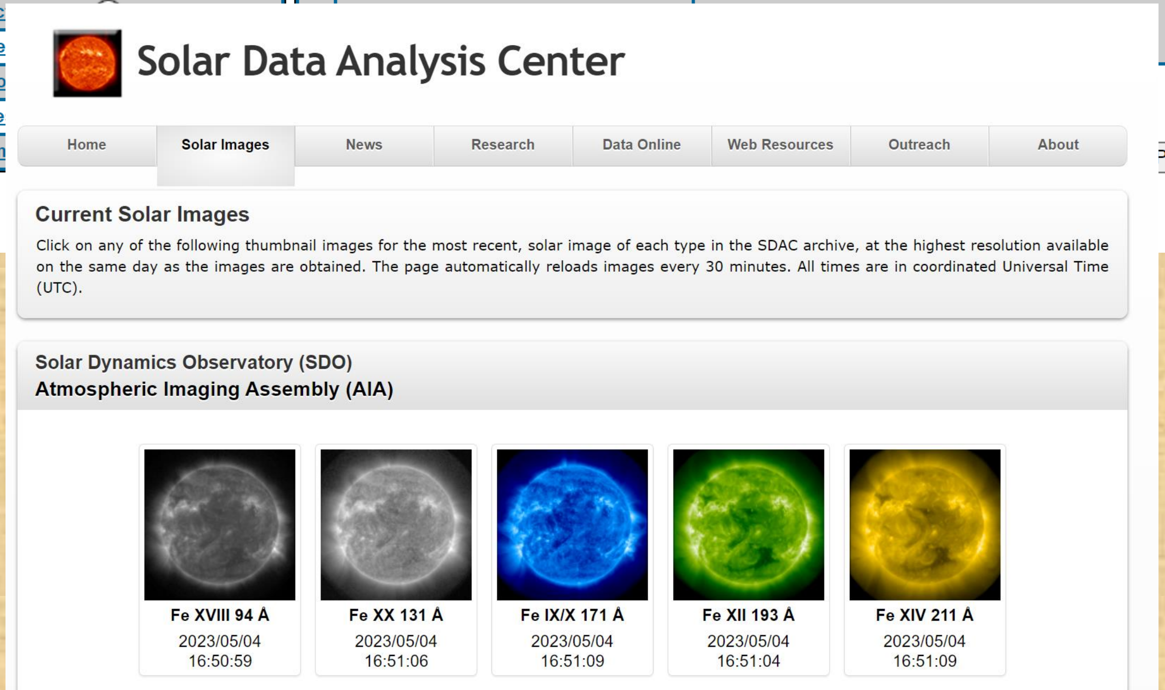
Current Product Restrictions — Remove All

Metadata contains 'electron'	Remove
Metadata contains 'energy'	Remove

Showing 301 - 320 of 562 Results — View Current List — Sort by Observatory

#	Products (& SPASE descriptions)	Information and Access Links
301	MMS 3 Electron Drift Instrument (EDI) Quality Zero Counts, Level 2 (L2), Survey Mode, 0.125 s Data https://doi.org/10.48322/cwb6-vf46	· FTPS from the MMS SDC (not with most browsers) · HTTPS from the MMS SDC · FTPS from SPDF (not with most browsers) · HTTPS from SPDF · CDAWeb · CDAWeb HAPI Server · The Magnetospheric Multiscale (MMS) Mission home page at Goddard Space Flight Center (GSFC) · Data Caveats and Current Release Notes at LASP MMS Get Data/Plots

Plots



Solar Data Analysis Center

Current Solar Images

Click on any of the following thumbnail images for the most recent, solar image of each type in the SDAC archive, at the highest resolution available on the same day as the images are obtained. The page automatically reloads images every 30 minutes. All times are in coordinated Universal Time (UTC).

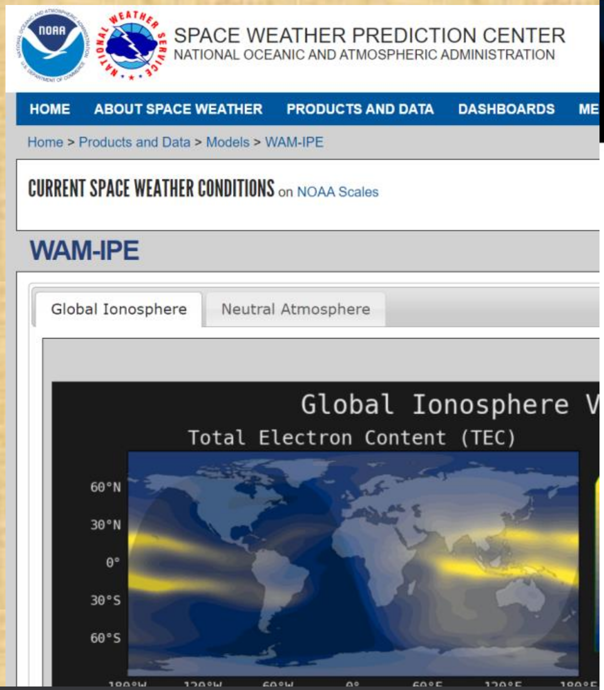
Solar Dynamics Observatory (SDO)
Atmospheric Imaging Assembly (AIA)

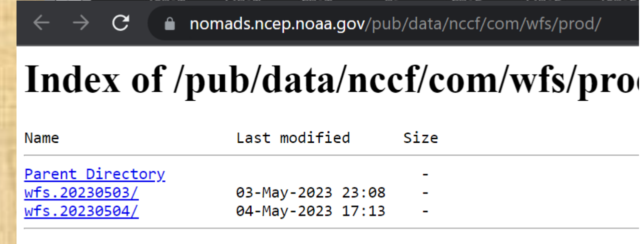
Fe XVIII 94 Å	Fe XX 131 Å	Fe IX/X 171 Å	Fe XII 193 Å	Fe XIV 211 Å
2023/05/04 16:50:59	2023/05/04 16:51:06	2023/05/04 16:51:09	2023/05/04 16:51:04	2023/05/04 16:51:09

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.

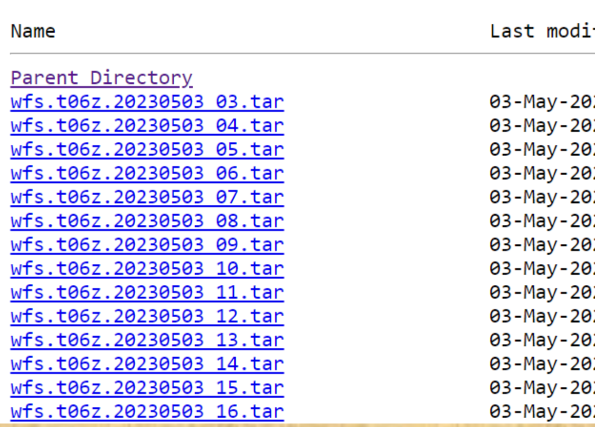
How can PyHC help with these issues?


SPACE WEATHER PREDICTION CENTER
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
HOME
ABOUT SPACE WEATHER
PRODUCTS AND DATA
DASHBOARDS
ME
Home > Products and Data > Models > WAM-IPE
CURRENT SPACE WEATHER CONDITIONS on NOAA Scales
WAM-IPE
Global Ionosphere
Neutral Atmosphere
Global Ionosphere V
Total Electron Content (TEC)
60°N
30°N
0°
30°S
60°S

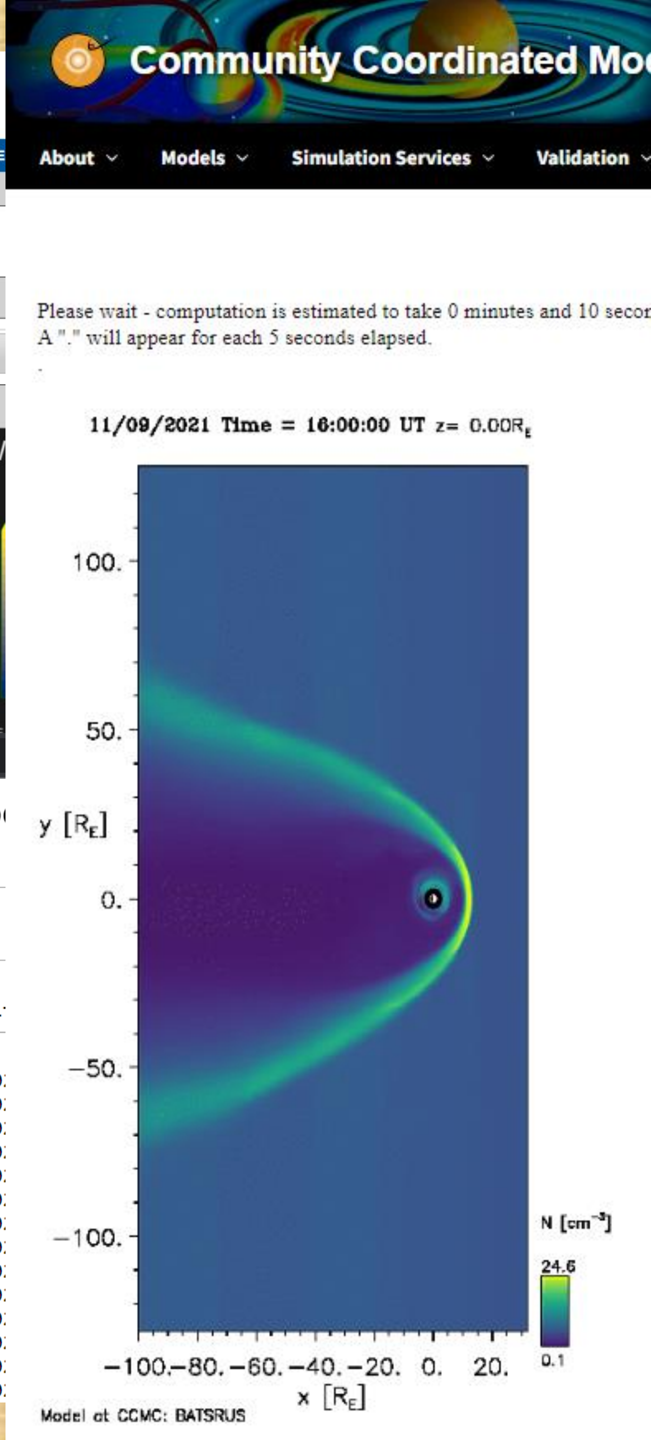


Index of /pub/data/nccf/com/wfs/pro

```
Name                Last modified      Size

Parent Directory                        -
wfs.20230503/       03-May-2023 23:08   -
wfs.20230504/       04-May-2023 17:13   -
```



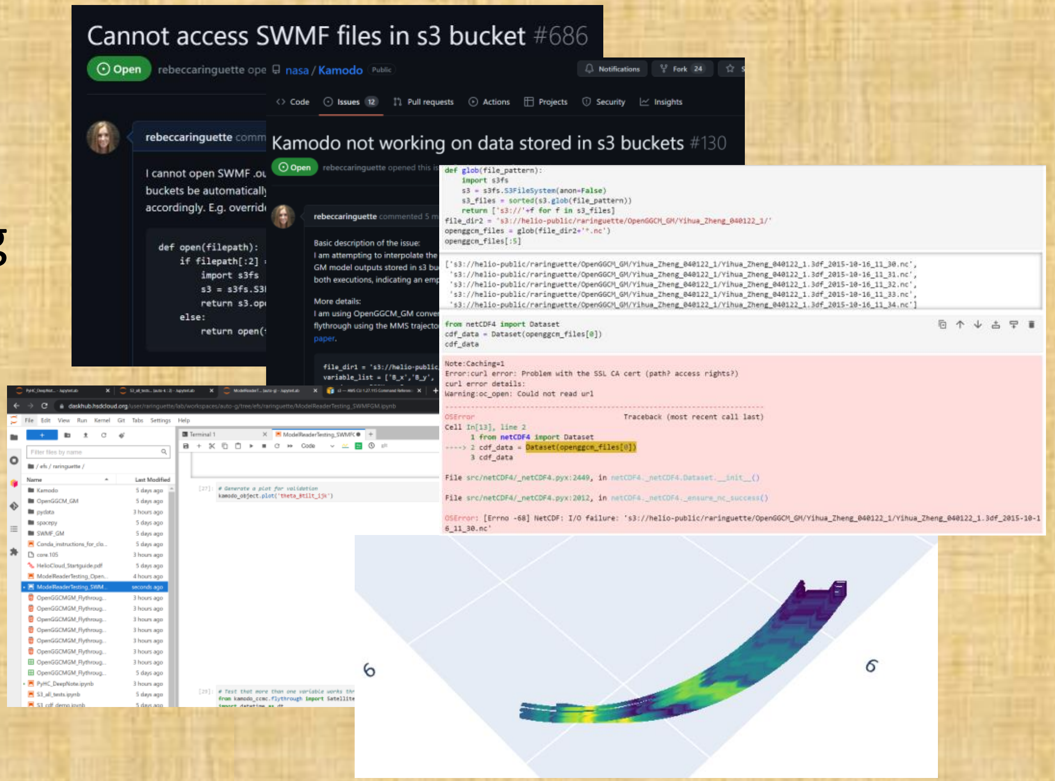
```
Name                                  Last modi

Parent Directory
wfs.t06z.20230503_03.tar              03-May-20
wfs.t06z.20230503_04.tar              03-May-20
wfs.t06z.20230503_05.tar              03-May-20
wfs.t06z.20230503_06.tar              03-May-20
wfs.t06z.20230503_07.tar              03-May-20
wfs.t06z.20230503_08.tar              03-May-20
wfs.t06z.20230503_09.tar              03-May-20
wfs.t06z.20230503_10.tar              03-May-20
wfs.t06z.20230503_11.tar              03-May-20
wfs.t06z.20230503_12.tar              03-May-20
wfs.t06z.20230503_13.tar              03-May-20
wfs.t06z.20230503_14.tar              03-May-20
wfs.t06z.20230503_15.tar              03-May-20
wfs.t06z.20230503_16.tar              03-May-20
```



Please wait - computation is estimated to take 0 minutes and 10 secon
A "." will appear for each 5 seconds elapsed.

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?



Cannot access SWMF files in s3 bucket #686

Open rebeccaringuette ope

rebeccaringuette comm

I cannot open SWMF .ou
buckets be automatically
accordingly. E.g. overrid

```
def open(filepath):
    if filepath[:2]
        import s3fs
        s3 = s3fs.S3
        return s3.op
    else:
        return open(
```

nasa / Kamodo Public

Kamodo not working on data stored in s3 buckets #130

Open rebeccaringuette opened this is

rebeccaringuette commented 5 m

Basic description of the issue:
I am attempting to interpolate the
GM model outputs stored in s3 bu
both executions, indicating an emp

More details:
I am using OpenGGCM_GM conver
flythrough using the MMS trajecto
paper.

```
file_dir1 = 's3://helio-public
variable_list = ['B_x','B_y',
```

```
def glob(file_pattern):
    import s3fs
    s3 = s3fs.S3FileSystem(anon=False)
    s3_files = sorted(s3.glob(file_pattern))
    return ['s3://'+f for f in s3_files]
file_dir2 = 's3://helio-public/raringuette/OpenGGCM_GM/Yihua_Zheng_040122_1/'
openggcm_files = glob(file_dir2+'*.nc')
openggcm_files[:5]
```

```
['s3://helio-public/raringuette/OpenGGCM_GM/Yihua_Zheng_040122_1/Yihua_Zheng_040122_1.3df_2015-10-16_11_30.nc',
 's3://helio-public/raringuette/OpenGGCM_GM/Yihua_Zheng_040122_1/Yihua_Zheng_040122_1.3df_2015-10-16_11_31.nc',
 's3://helio-public/raringuette/OpenGGCM_GM/Yihua_Zheng_040122_1/Yihua_Zheng_040122_1.3df_2015-10-16_11_32.nc',
 's3://helio-public/raringuette/OpenGGCM_GM/Yihua_Zheng_040122_1/Yihua_Zheng_040122_1.3df_2015-10-16_11_33.nc',
 's3://helio-public/raringuette/OpenGGCM_GM/Yihua_Zheng_040122_1/Yihua_Zheng_040122_1.3df_2015-10-16_11_34.nc']
```

```
from netCDF4 import Dataset
cdf_data = Dataset(openggcm_files[0])
cdf_data
```

```
Note:Caching=1
Error:curl error: Problem with the SSL CA cert (path? access rights?)
curl error details:
Warning:oc_open: Could not read url
---------------------------------------------------------------------------
OSError                                   Traceback (most recent call last)
Cell In[13], line 2
      1 from netCDF4 import Dataset
----> 2 cdf_data = Dataset(openggcm_files[0])
      3 cdf_data

File src/netCDF4/_netCDF4.pyx:2449, in netCDF4._netCDF4.Dataset.__init__()

File src/netCDF4/_netCDF4.pyx:2012, in netCDF4._netCDF4._ensure_nc_success()

OSError: [Errno -68] NetCDF: I/O failure: 's3://helio-public/raringuette/OpenGGCM_GM/Yihua_Zheng_040122_1/Yihua_Zheng_040122_1.3df_2015-10-16_11_30.nc'
```

daskhub.hsdcloud.org/user/raringuette/lab/workspaces/auto-g/tree/efs/raringuette/ModelReaderTesting_SWMFGM.ipynb

PyHC should take the lead here. What paths forward have low-hanging fruit?

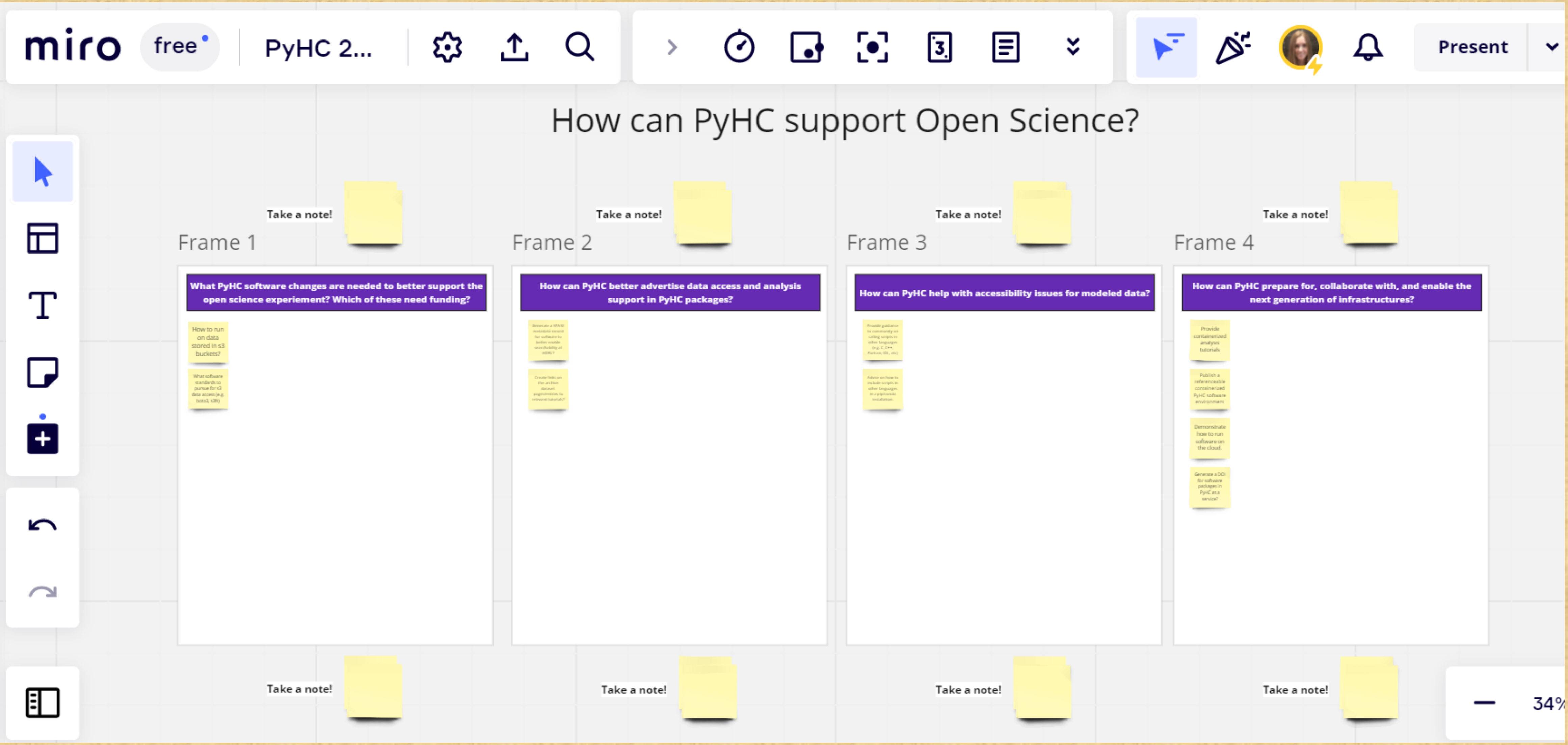
Where is this going?

- 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



How can PyHC support Open Science?

Take a note!

Frame 1

What PyHC software changes are needed to better support the open science experiement? Which of these need funding?

How to run on data stored in s3 buckets?

What software standards to pursue for s3 data access (e.g. boto3, s3fs)

Take a note!

Take a note!

Frame 2

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

Generate a SPASE metadata record for software to better enable searchability at HDRL?

Create links on the archive dataset pages/entities to relevant tutorials?

Take a note!

Take a note!

Frame 3

How can PyHC help with accessibility issues for modeled data?

Provide guidance to community on calling scripts in other languages (e.g. C, C++, Fortran, IDL, etc.)

Advise on how to include scripts in other languages in a pip/conda installation.

Take a note!

Take a note!

Frame 4

How can PyHC prepare for, collaborate with, and enable the next generation of infrastructures?

Provide containerized analysis tutorials

Publish a referenceable containerized PyHC software environment

Demonstrate how to run software on the cloud.

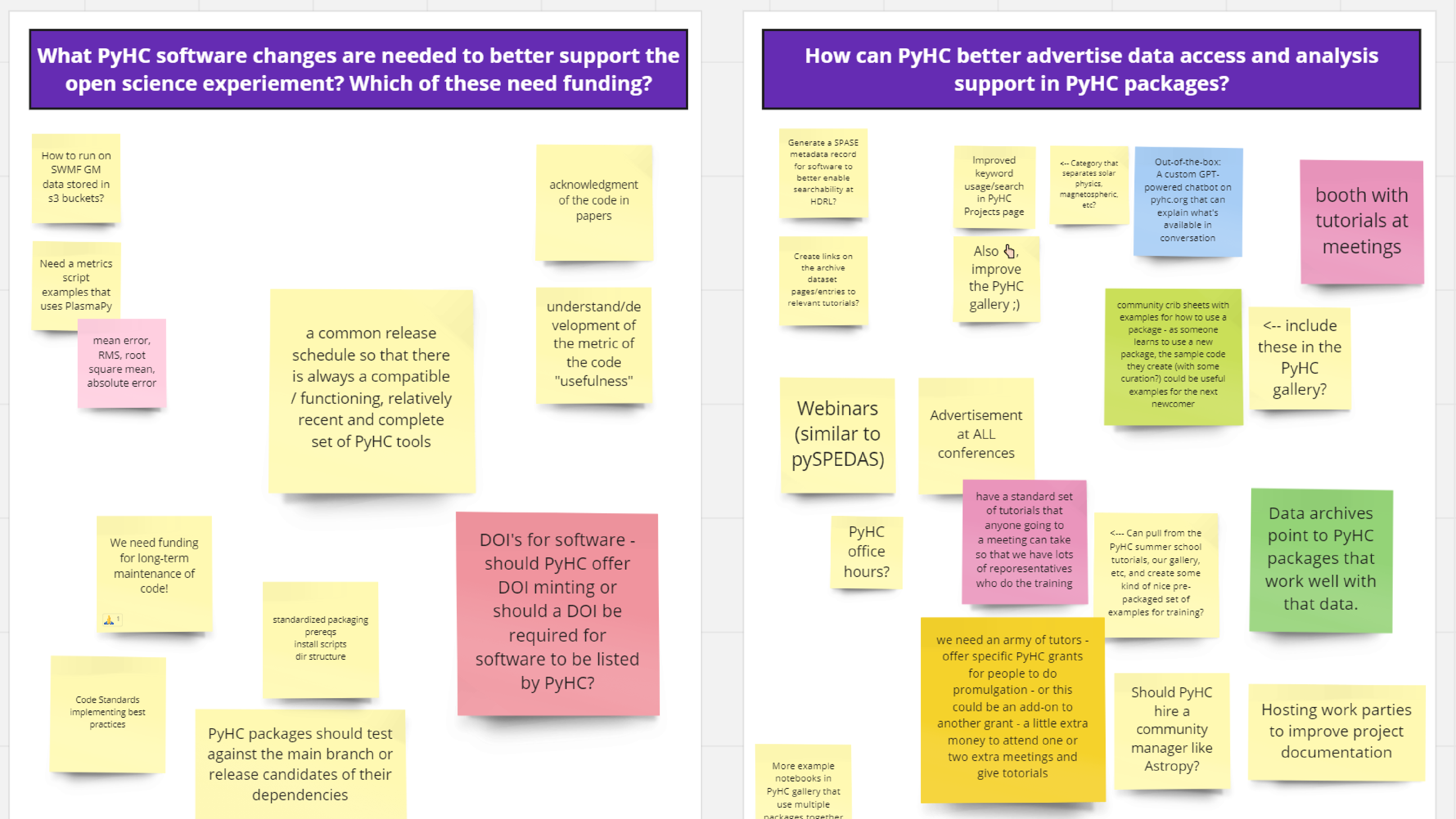
Generate a DOI for software packages in PyHC as a service?

Take a note!

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 experiement? Which of these need funding?

How to run on SWMF GM data stored in s3 buckets?

Need a metrics script examples that uses PlasmaPy

mean error, RMS, root square mean, absolute error

acknowledgment of the code in papers

understand/development of the metric of the code "usefulness"

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!

standardized packaging
prereqs
install scripts
dir structure

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

Code Standards implementing best practices

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

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

Generate a SPASE metadata record for software to better enable searchability at HDRL?

Create links on the archive dataset pages/entries to relevant tutorials?

Improved keyword usage/search in PyHC Projects page

Also 👆, improve the PyHC gallery ;)

<-- Category that separates solar physics, magnetospheric, etc?

Out-of-the-box: A custom GPT-powered chatbot on pyhc.org that can explain what's available in conversation

booth with tutorials at meetings

community crib sheets with examples for how to use a package - as someone learns to use a new package, the sample code they create (with some curation?) could be useful examples for the next newcomer

<-- include these in the PyHC gallery?

Webinars (similar to pySPEDAS)

Advertisement at ALL conferences

PyHC office hours?

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

<--- Can pull from the PyHC summer school tutorials, our gallery, etc, and create some kind of nice pre-packaged set of examples for training?

Data archives point to PyHC packages that work well with that data.

we need an army of tutors - offer specific PyHC grants for people to do promulgation - or this could be an add-on to another grant - a little extra money to attend one or two extra meetings and give totorials

Should PyHC hire a community manager like Astropy?

Hosting work parties to improve project documentation

More example notebooks in PyHC gallery that use multiple packages together

How can PyHC help with accessibility issues for modeled data?

Provide guidance to community on calling scripts in other languages (e.g. C, C++, Fortran, IDL, etc)

Advise on how to include scripts in other languages as dependencies in a pip/conda installation.

Simplify the process to have SpacePy as a dependency

Create metadata standard(s) for simulation inputs/outputs

standardized Data Storage format
access methods

get CCMC to engage more with the community and find out more about what users need

modeling results hosting/server should exist

many modelers are using Paraview, so PyHC could spend resources to support Paraview

make a specific call in the HTM NASA AO for improving model access

Partner with IHDEA

How can PyHC prepare for, collaborate with, and enable the next generation of infrastructures?

Provide containerized analysis tutorials

Publish a referenceable containerized PyHC software environment

Demonstrate how to run software on the cloud.

Generate a DOI for software packages in PyHC as a service

What software standards to pursue for s3 data access (e.g. boto3, s3fs)?

data access beyond local files of all kinds (HAPI, other companies' clouds)

Improve culture by building a foundation of psychological safety

Per Gretche's diagram where the last step before making open science required is making open science being rewarding, we should advertise how it's rewarding (and if we can't figure out how its rewarding to individuals, we reconsider making it more rewarding)

is Podman/Docker a viable 'package system' for prereqs+code to work?

https://docs.sylabs.io/guides/3.5/user-guide/introduction.html

(singularity)

Continued outreach to the next generation of scientific software developers and scientists through engaging and informative PyHC Summer schools, presentations and booths at other meetings, etc.

Funding for a software engineer who focuses on package/docs/testing infrastructure?

Docker alternative: https://podman.io/

The Execution Model prevent arbitrary code execution (aka no bitcoin mining and spam)

yearly simultaneous release of all PyHC packages for compatibility

Hackathon or similar to try out a lot of different ways to containerize/compartmentalize a turnkey py?HC environment

allow multiple options for obtaining a viable collection of PyHC tooling

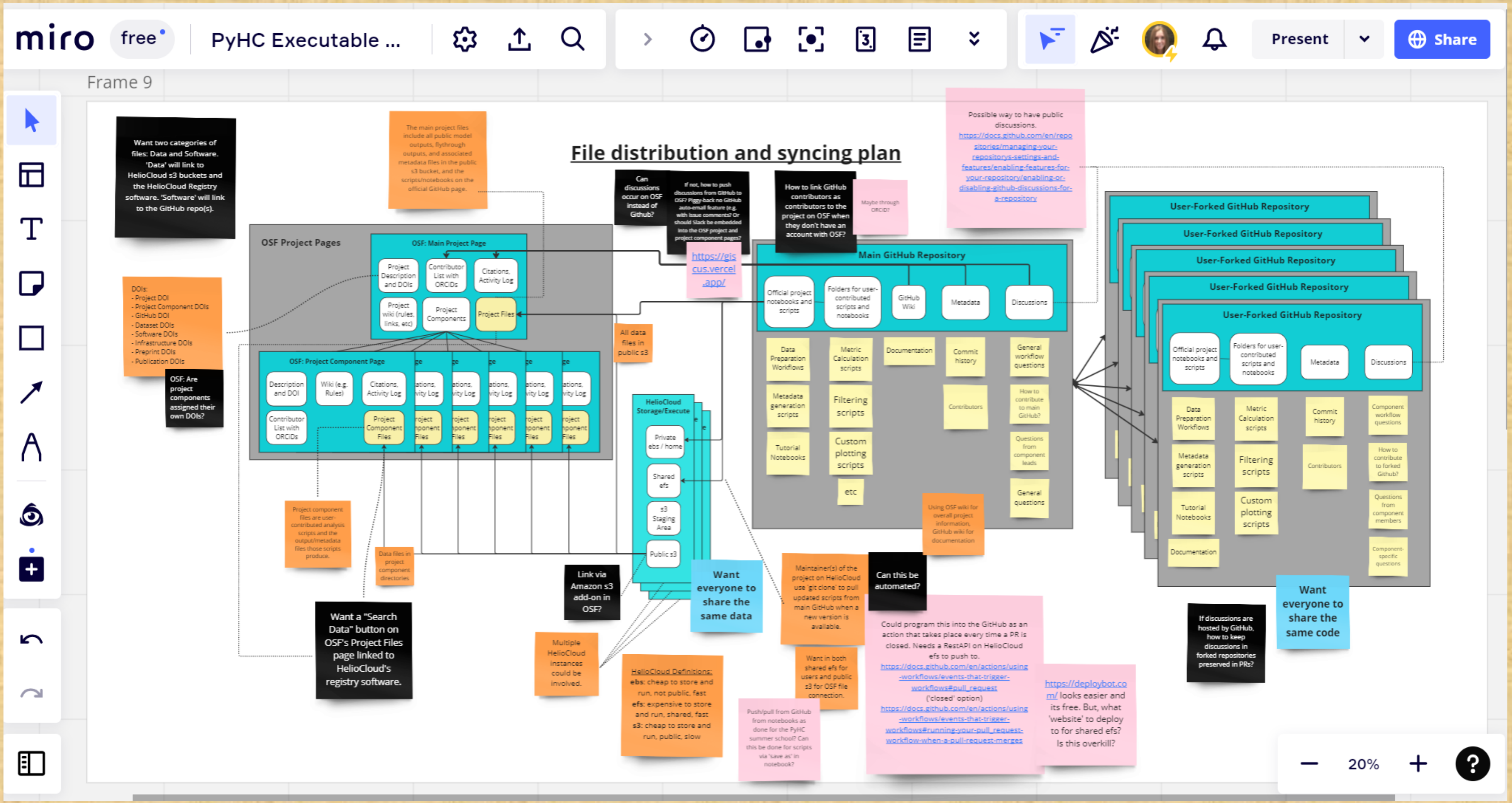
use singularity instead of Docker

Hardware dependence/support (large data, high core count, large memory, GPUs)

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-c3a5772e5ce24ce1942b001696d52251
- This presentation’s link on PyHC’s google drive: https://docs.google.com/presentation/d/1c2bP0zDdiJWMCPZZxzm9U3zx80NPH_SC/edit?usp=share_link&ouid=118198339287841207428&rtpof=true&sd=true
- Miro board link with preliminary project workflows (view): https://miro.com/welcomeonboard/Q05NeGI2M0taVGtxekJjZXkzVzJsdzFud3R3SlF4RWJ4RGN0NXBmazZheThhd1d4aUNLclVDOFM1WHhXa01ZWXwzMDc0NDU3MzU3OTk0ODcyOTEyfDI=?share_link_id=953968385486
- Miro board link with PyHC 2023 spring session discussion (edit): https://miro.com/welcomeonboard/M2JRd2NNSXhPcXpPRXZJbDFHczcwRHNqYzUzSk9XelBzeVdkT2p3NThhQlVBRkQ4VEJzaXc5MWNGT3dhUklxR3wzMDc0NDU3MzU3OTk0ODcyOTEyfDI=?share_link_id=522570131865

Project Tour: Preliminary Project Workflows


miro
free
PyHC Executable ...
Present
Share
Frame 9
20%
OSF Project Pages
OSF: Main Project Page
Project Description and DOIs
Contributor List with ORCiDs
Citations, Activity Log
Project wiki (rules, links, etc)
Project Components
Project Files
OSF: Project Component Page
Description and DOI
Wiki (e.g. Rules)
Citations, Activity Log
Contributor List with ORCiDs
Project Component Files
HelioCloud Storage/Execute
Private ebs / home
Shared efs
s3 Staging Area
Public s3
Main GitHub Repository
Official project notebooks and scripts
Folders for user-contributed scripts and notebooks
GitHub Wiki
Metadata
Discussions
Data Preparation Workflows
Metadata generation scripts
Tutorial Notebooks
Metric Calculation scripts
Filtering scripts
Custom plotting scripts
etc
Documentation
Commit history
Contributors
General workflow questions
How to contribute to main GitHub?
Questions from component leads
General questions
User-Forked GitHub Repository
Official project notebooks and scripts
Folders for user-contributed scripts and notebooks
Metadata
Discussions
Data Preparation Workflows
Metadata generation scripts
Tutorial Notebooks
Documentation
Metric Calculation scripts
Filtering scripts
Custom plotting scripts
Commit history
Contributors
Component workflow questions
How to contribute to forked Github?
Questions from component members
Component-specific questions

File distribution and syncing plan

Want two categories of files: Data and Software. 'Data' will link to HelioCloud s3 buckets and the HelioCloud Registry software. 'Software' will link to the GitHub repo(s).

The main project files include all public model outputs, flythrough outputs, and associated metadata files in the public s3 bucket, and the scripts/notebooks on the official GitHub page.

DOIs:
- Project DOI
- Project Component DOIs
- GitHub DOI
- Dataset DOIs
- Software DOIs
- Infrastructure DOIs
- Preprint DOIs
- Publication DOIs

OSF: Are project components assigned their own DOIs?

Can discussions occur on OSF instead of Github?

If not, how to push discussions from GitHub to OSF? Piggy-back no GitHub auto-email feature (e.g. with issue comments? Or should Slack be embedded into the OSF project and project component pages?

https://giscus.vercel.app/

How to link GitHub contributors as contributors to the project on OSF when they don't have an account with OSF?

Maybe through ORCID?

Possible way to have public discussions. https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/enabling-features-for-your-repository/enabling-or-disabling-github-discussions-for-a-repository

All data files in public s3

Project component files are user-contributed analysis scripts and the output/metadata files those scripts produce.

Data files in project component directories

Want a "Search Data" button on OSF's Project Files page linked to HelioCloud's registry software.

Link via Amazon s3 add-on in OSF?

Multiple HelioCloud instances could be involved.

HelioCloud Definitions:
ebs: cheap to store and run, not public, fast
efs: expensive to store and run, shared, fast
s3: cheap to store and run, public, slow

Want everyone to share the same data

Maintainer(s) of the project on HelioCloud use 'git clone' to pull updated scripts from main GitHub when a new version is available.

Want in both shared efs for users and public s3 for OSF file connection.

Push/pull from GitHub from notebooks as done for the PyHC summer school? Can this be done for scripts via 'save as' in notebook?

Using OSF wiki for overall project information, GitHub wiki for documentation

Can this be automated?

Could program this into the GitHub as an action that takes place every time a PR is closed. Needs a RestAPI on HelioCloud efs to push to. https://docs.github.com/en/actions/using-workflows/events-that-trigger-workflows#pull_request ('closed' option) https://docs.github.com/en/actions/using-workflows/events-that-trigger-workflows#running-your-pull_request-workflow-when-a-pull-request-merges

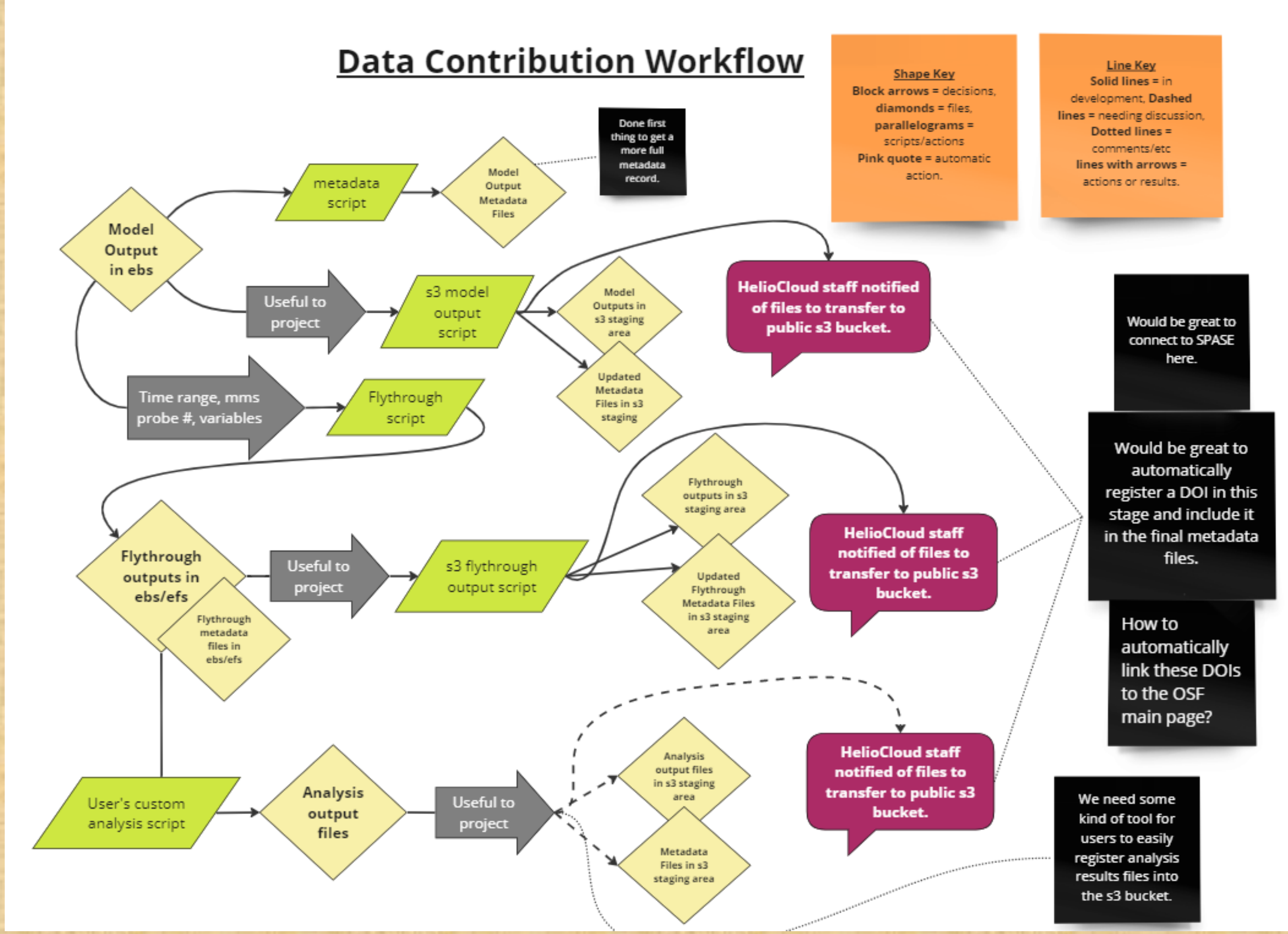
https://deploybot.com/ looks easier and its free. But, what 'website' to deploy to for shared efs? Is this overkill?

If discussions are hosted by GitHub, how to keep discussions in forked repositories preserved in PRs?

Want everyone to share the same code

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

Project Tour: Preliminary Project Workflows



Data Contribution Workflow

Shape Key
Block arrows = decisions, **diamonds =** files, **parallelograms =** scripts/actions
Pink quote = automatic action.

Line Key
Solid lines = in development, **Dashed lines =** needing discussion, **Dotted lines =** comments/etc
lines with arrows = actions or results.

Model Output in ebs

metadata script

Model Output Metadata Files

Done first thing to get a more full metadata record.

Useful to project

s3 model output script

Model Outputs in s3 staging area

Updated Metadata Files in s3 staging

HelioCloud staff notified of files to transfer to public s3 bucket.

Time range, mms probe #, variables

Flythrough script

Flythrough outputs in ebs/efs

Flythrough metadata files in ebs/efs

Useful to project

s3 flythrough output script

Flythrough outputs in s3 staging area

Updated Flythrough Metadata Files in s3 staging area

HelioCloud staff notified of files to transfer to public s3 bucket.

User's custom analysis script

Analysis output files

Useful to project

Analysis output files in s3 staging area

Metadata Files in s3 staging area

HelioCloud staff notified of files to transfer to public s3 bucket.

Would be great to connect to SPASE here.

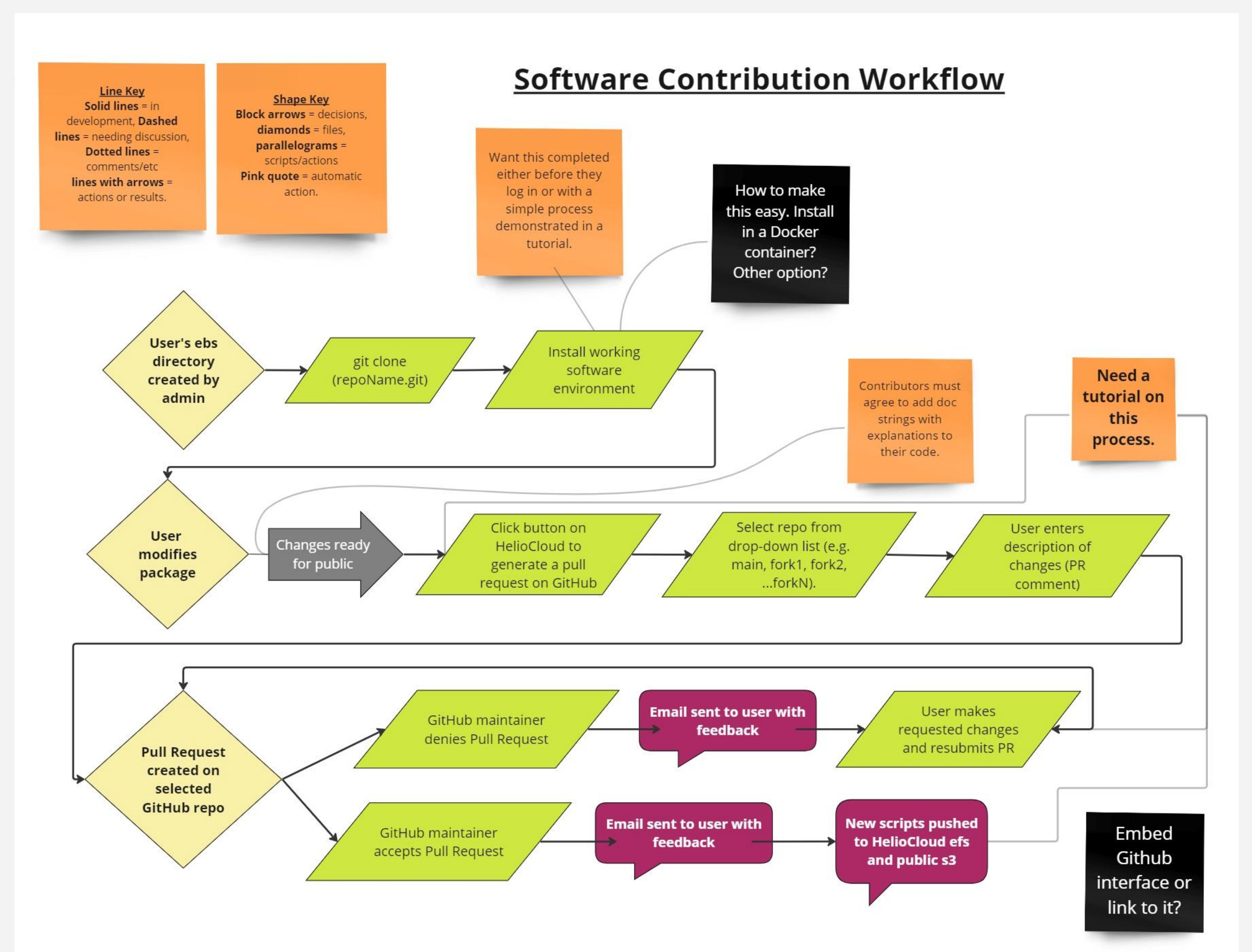
Would be great to automatically register a DOI in this stage and include it in the final metadata files.

How to automatically link these DOIs to the OSF main page?

We need some kind of tool for users to easily register analysis results files into the s3 bucket.

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

Project Tour: Preliminary Project Workflows



Software Contribution Workflow

Line Key
Solid lines = in development, **Dashed lines** = needing discussion, **Dotted lines** = comments/etc
lines with arrows = actions or results.

Shape Key
Block arrows = decisions, **diamonds** = files, **parallelograms** = scripts/actions
Pink quote = automatic action.

Want this completed either before they log in or with a simple process demonstrated in a tutorial.

How to make this easy. Install in a Docker container? Other option?

User's ebs directory created by admin

git clone (repoName.git)

Install working software environment

Contributors must agree to add doc strings with explanations to their code.

Need a tutorial on this process.

User modifies package

Changes ready for public

Click button on HelioCloud to generate a pull request on GitHub

Select repo from drop-down list (e.g. main, fork1, fork2, ...forkN).

User enters description of changes (PR comment)

Pull Request created on selected GitHub repo

GitHub maintainer denies Pull Request

Email sent to user with feedback

User makes requested changes and resubmits PR

GitHub maintainer accepts Pull Request

Email sent to user with feedback

New scripts pushed to HelioCloud efs and public s3

Embed Github interface or link to it?

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