

# Research Infrastructure Needs for Collaborative Science

Dr. Rebecca Ringuette (and many others)

NASA GSFC Center for HelioAnalytics

ISTPNext Workshop: May 8-10, 2023 at JHU/APL

# Drivers of Infrastructure Development

- Observation coordination across missions
- Multi-mission science analyses
- Big Data (observed and modeled)
- AI/ML data input requirements
- Lack of computational power
- Cross-domain and cross-disciplinary collaborations
- Executable papers
- Collaborative software analysis environments

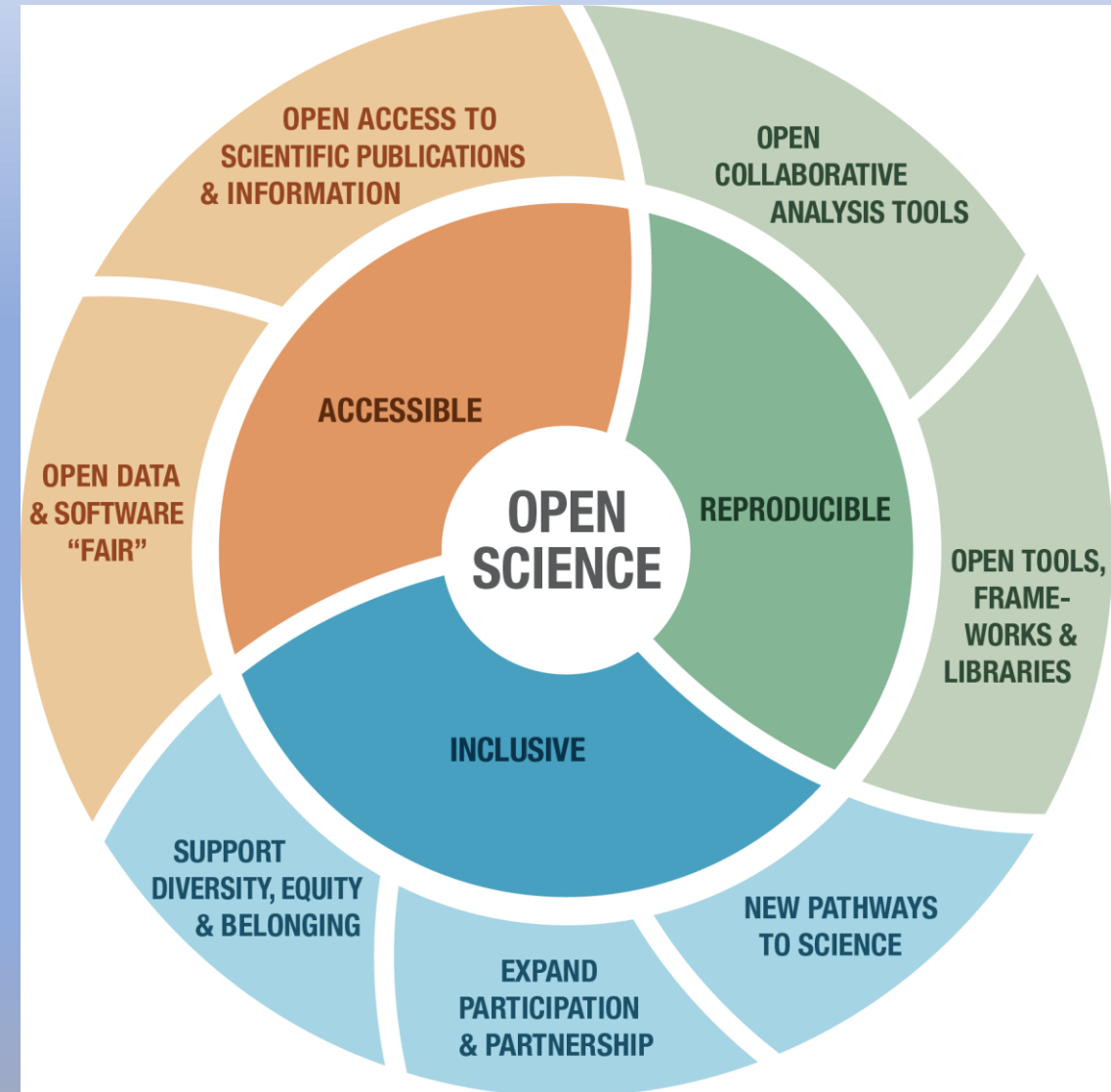
***All of these efforts can be accelerated by Open Science.***

# What is Open Science?

- ***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>)



# What Does Open Science Require?


- **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*: models, software environments,
  - *The great unknown*: people, relationships, collaborations, ...
- **Reproducible** results
  - Executable papers? 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**.
- Many observational datasets still lack these properties, so **continued progress is needed**.

**GODDARD SPACE FLIGHT CENTER**  
Space Physics Data Facility

+ Goddard Home  
+ Visit NASA.gov

**Heliophysics Data Portal** “Find it. Browse it. Get it.”

**SPASE**  
inside

HelpGeo OrbitsHelio OrbitsSPASE RegistryADS AbstractsFeedback

**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

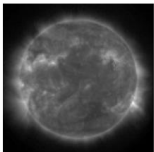
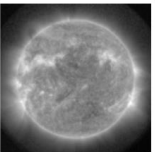
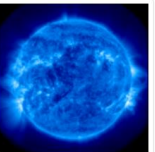
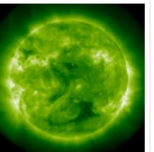
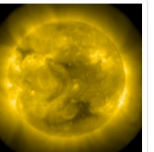
#	Products (& SPASE descriptions)	Information and Access Links
301	MMS 3 Electron Drift Instrument (EDI) Quality Zero Counts, Level 2 (L2), Survey Mode, 0.125 s Data  <a href="https://doi.org/10.48322/cwb6-vf46">https://doi.org/10.48322/cwb6-vf46</a>	<ul style="list-style-type: none"><li>• FTPS from the MMS SDC (not with most browsers)</li><li>• HTTPS from the MMS SDC</li><li>• FTPS from SPDF (not with most browsers)</li><li>• HTTPS from SPDF</li><li>• CDAWeb</li><li>• CDAWeb HAPI Server</li><li>• The Magnetospheric Multiscale (MMS) Mission home page at Goddard Space Flight Center (GSFC)</li><li>• Data Caveats and Current Release Notes at LASP MMS</li></ul> <a href="#">Get Data/Plots</a>

**Solar Data Analysis Center**

HomeSolar ImagesNewsResearchData OnlineWeb ResourcesOutreachAbout

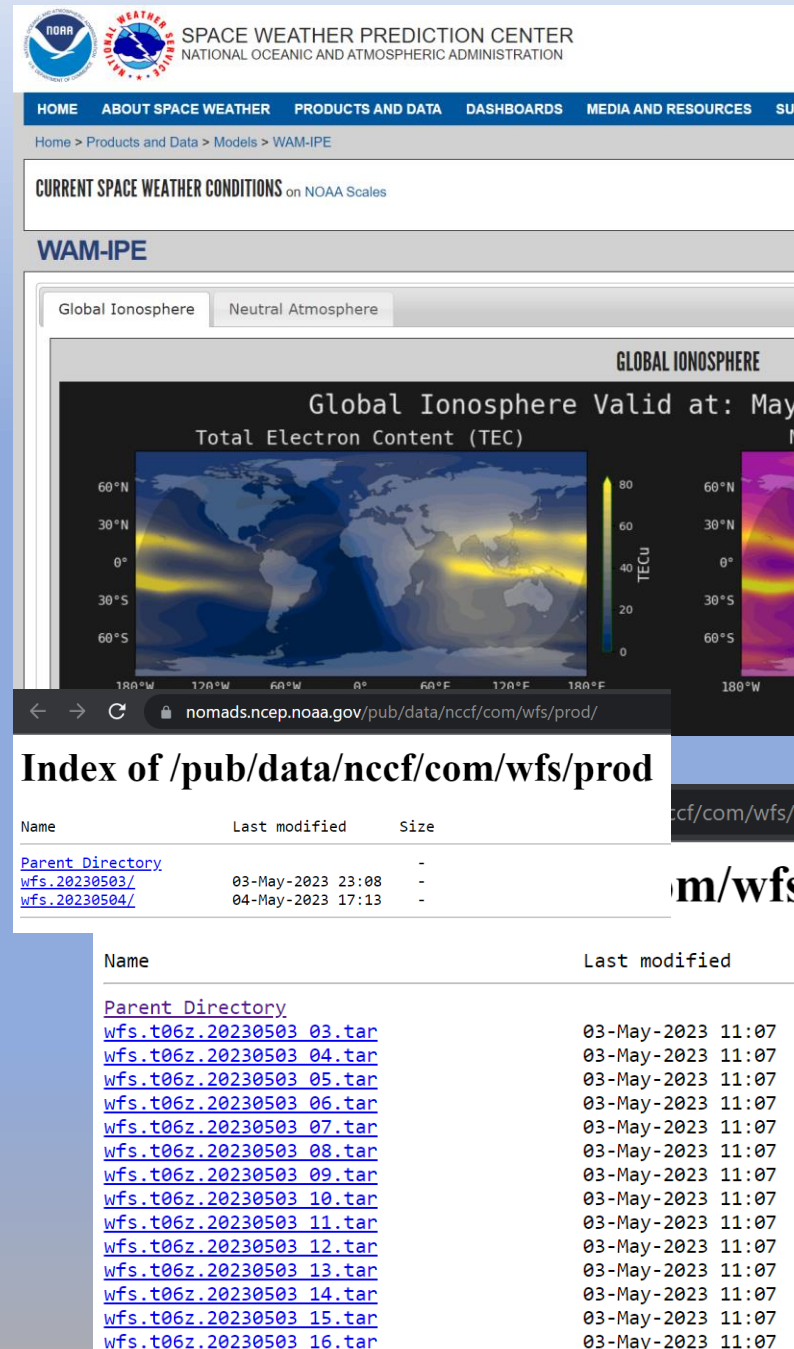
**Current Solar Images**  
Click on any of the following thumbnail images for the most recent, solar image of each type in the SDAC archive, at the highest resolution available on the same day as the images are obtained. The page automatically reloads images every 30 minutes. All times are in coordinated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

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

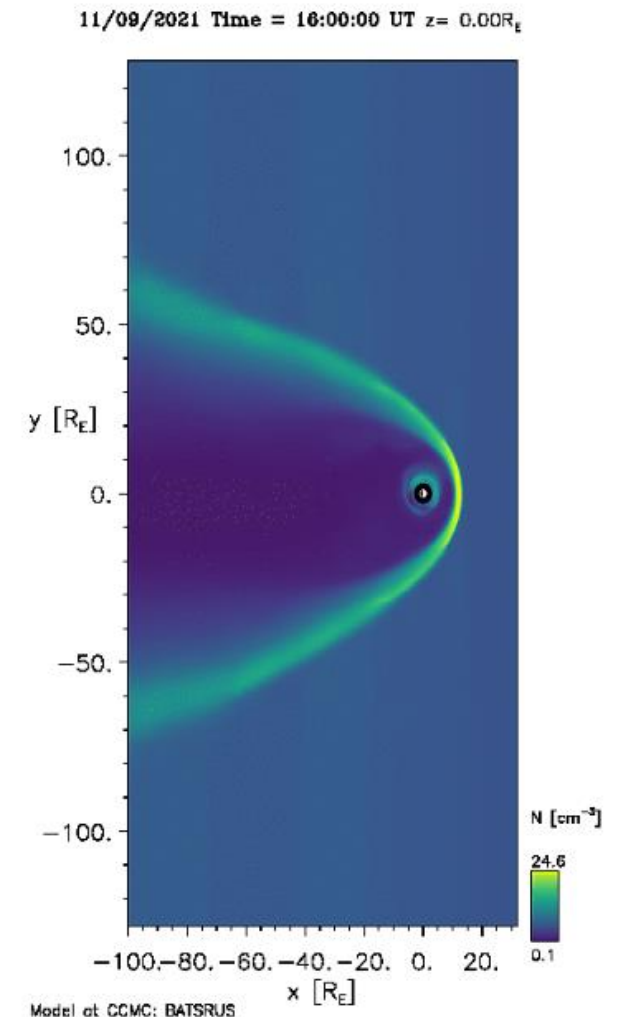
				
<b>Fe XVIII 94 Å</b> 2023/05/04 16:50:59	<b>Fe XX 131 Å</b> 2023/05/04 16:51:06	<b>Fe IX/X 171 Å</b> 2023/05/04 16:51:09	<b>Fe XII 193 Å</b> 2023/05/04 16:51:04	<b>Fe XIV 211 Å</b> 2023/05/04 16:51:09

# Modeled Data

- Infrastructure supporting modeled data is **far less developed**.
  - No modern **search** interfaces are known,
  - **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.

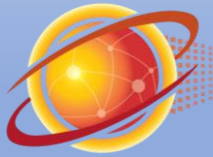


Please wait - computation is estimated to take 0 minutes and 10 seconds. A "." will appear for each 5 seconds elapsed.





# Modeled Data



HelioCloud

- Modeled data are more **voluminous** than observational data, ***complicating the accessibility problem.***
  - Need an online/hosted analysis platform (e.g. HelioCloud) for users to down-select and filter modeled data to reduce downloaded volumes.
- Modeled data is (somewhat) **reproducible, but only IF:**
  - Containerization instructions and full run instructions are shared,
  - Containerized model codes are shared with an open license, and
  - All run inputs, settings and related information are shared.
- Modeled data's **metadata** are lacking compared to observed data.
  - Must indicate generating model, version number, model's DOI, all run settings, input data sources, DOI for the specific containerized model instance used, containerization method instructions, generating institution, computing infrastructure description, run name/ID, and more.
  - Must also include start and stop times, variables contained, coordinate systems, coordinate ranges, domains, and more.

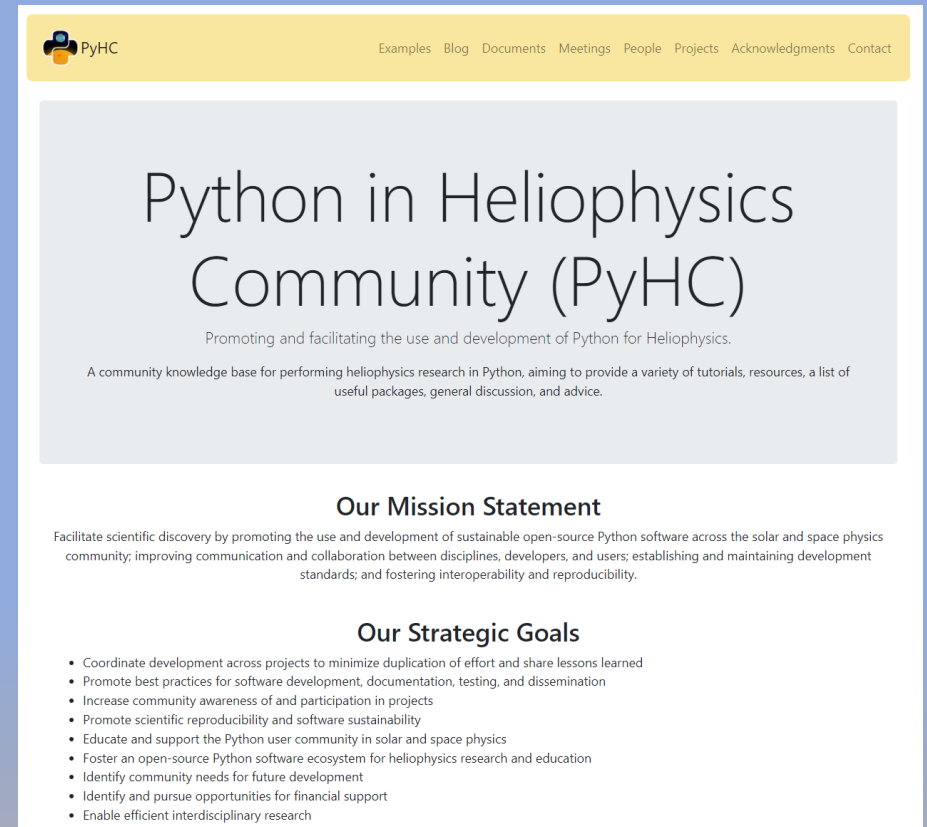
Reproducibility

Selection

# Software

- Sustained push is needed 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 is currently working out these issues.

<https://heliopython.org/>



The screenshot shows the homepage of the Python in Heliophysics Community (PyHC). The header is yellow with the PyHC logo and a navigation menu: Examples, Blog, Documents, Meetings, People, Projects, Acknowledgments, Contact. The main content area has a light gray background with the title "Python in Heliophysics Community (PyHC)" in large black font. Below the title is a subtitle: "Promoting and facilitating the use and development of Python for Heliophysics." and a paragraph: "A community knowledge base for performing heliophysics research in Python, aiming to provide a variety of tutorials, resources, a list of useful packages, general discussion, and advice." Below this is a section titled "Our Mission Statement" with a paragraph: "Facilitate scientific discovery by promoting the use and development of sustainable open-source Python software across the solar and space physics community; improving communication and collaboration between disciplines, developers, and users; establishing and maintaining development standards; and fostering interoperability and reproducibility." Below that is a section titled "Our Strategic Goals" with a bulleted list: "Coordinate development across projects to minimize duplication of effort and share lessons learned", "Promote best practices for software development, documentation, testing, and dissemination", "Increase community awareness of and participation in projects", "Promote scientific reproducibility and software sustainability", "Educate and support the Python user community in solar and space physics", "Foster an open-source Python software ecosystem for heliophysics research and education", "Identify community needs for future development", "Identify and pursue opportunities for financial support", and "Enable efficient interdisciplinary research".



# Infrastructure Development Plan

- Continue the **push towards FAIR observed data** (both space physics and solar physics).
- Determine the infrastructure and standards needed to make **software FAIR**.
- **Focus primarily** on developing infrastructure for **modeled data**:
  - **Searchable** through a modern web-based interface built on SPASE,
  - **Downloadable** from webpages and through APIs,
  - **Citable** with DOIs (independent of publications), and
  - Ideally available in file formats compatible with **efficient computation on the cloud** (e.g. netCDF4/HDF5 as Zarr),



# How will we use it?



- 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).

*Open science demands a significant investment into infrastructure development.*