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

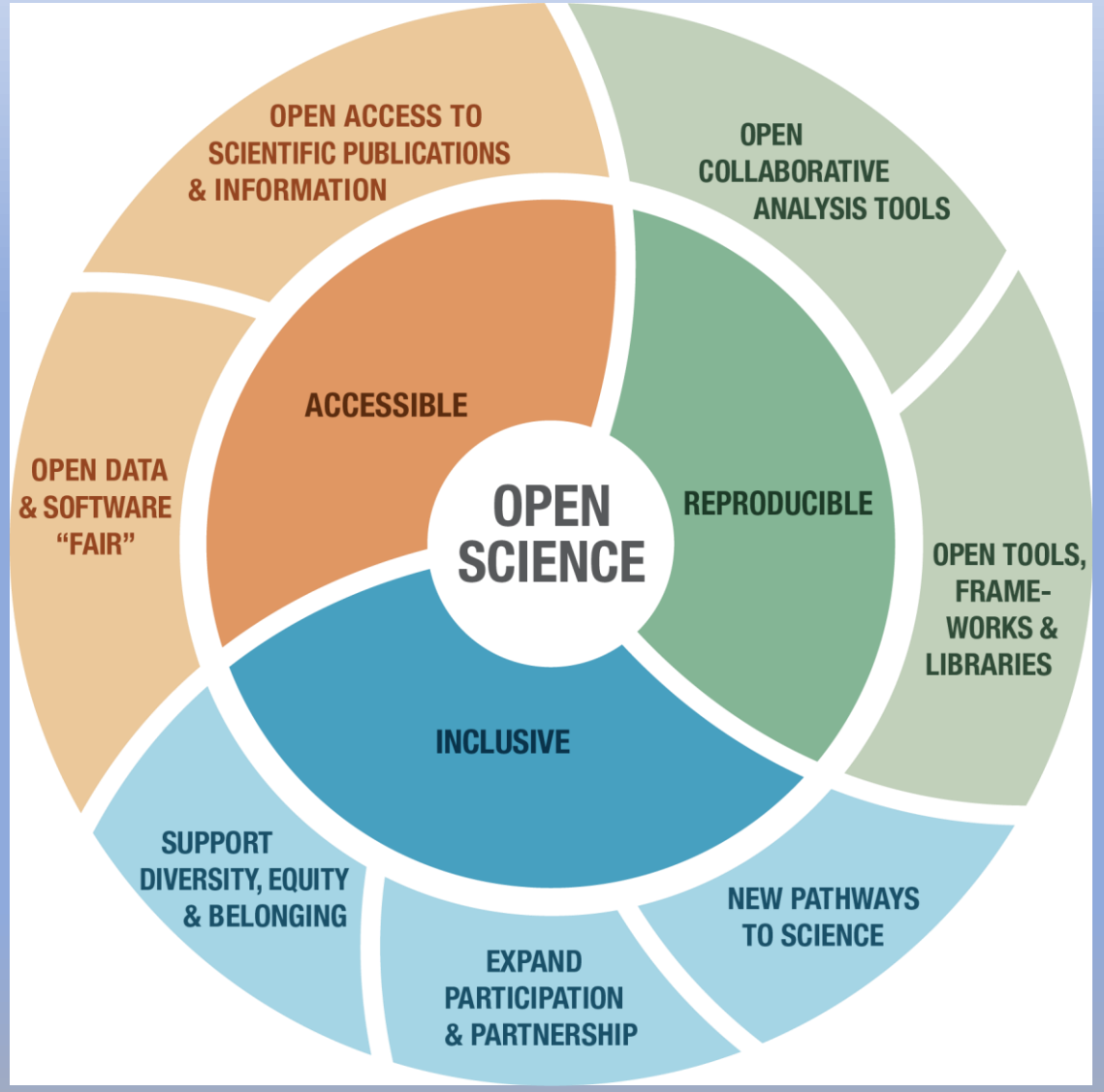
Drivers of Infrastructure Development

- 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

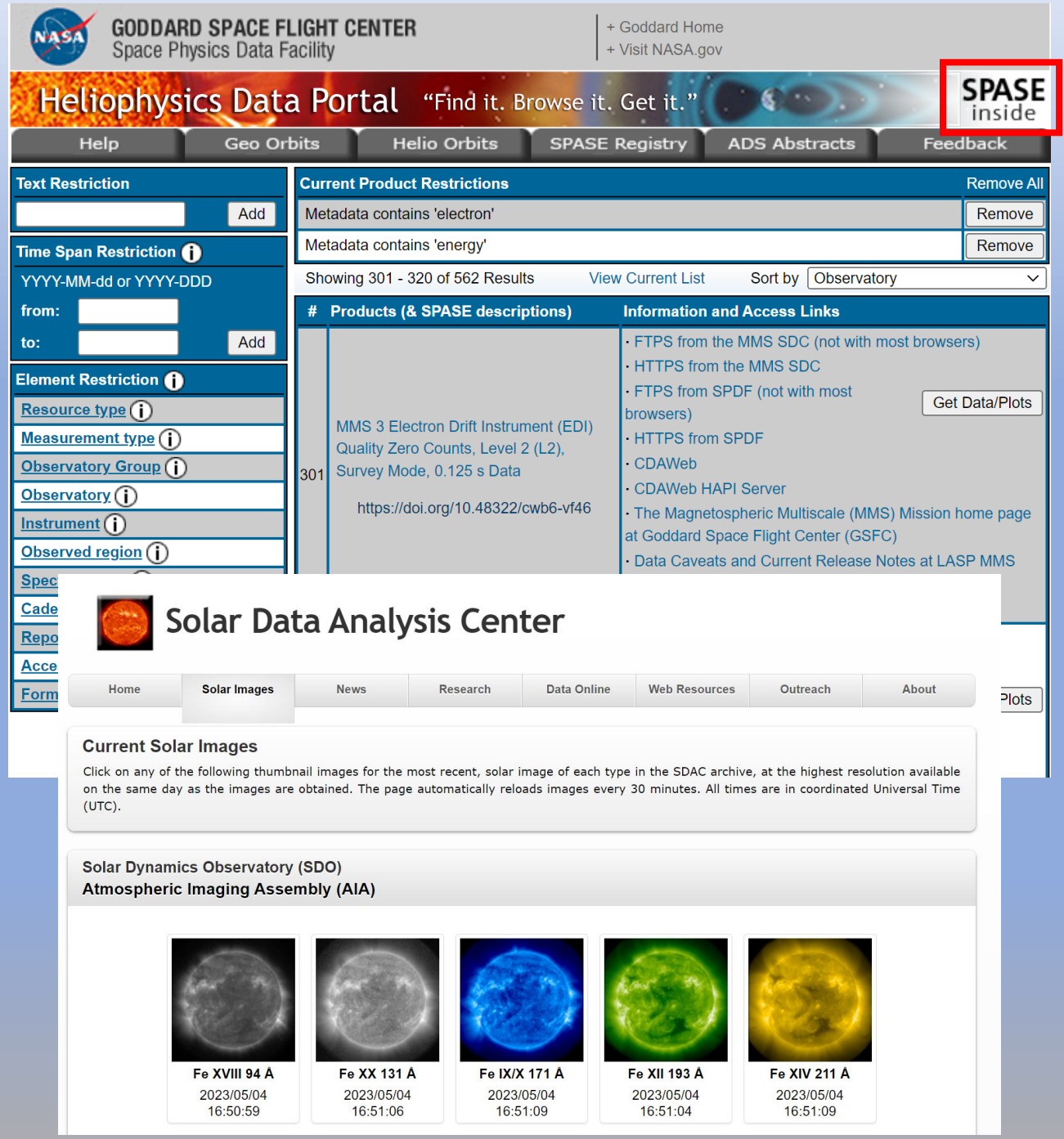
  How to make collaborations open?

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.

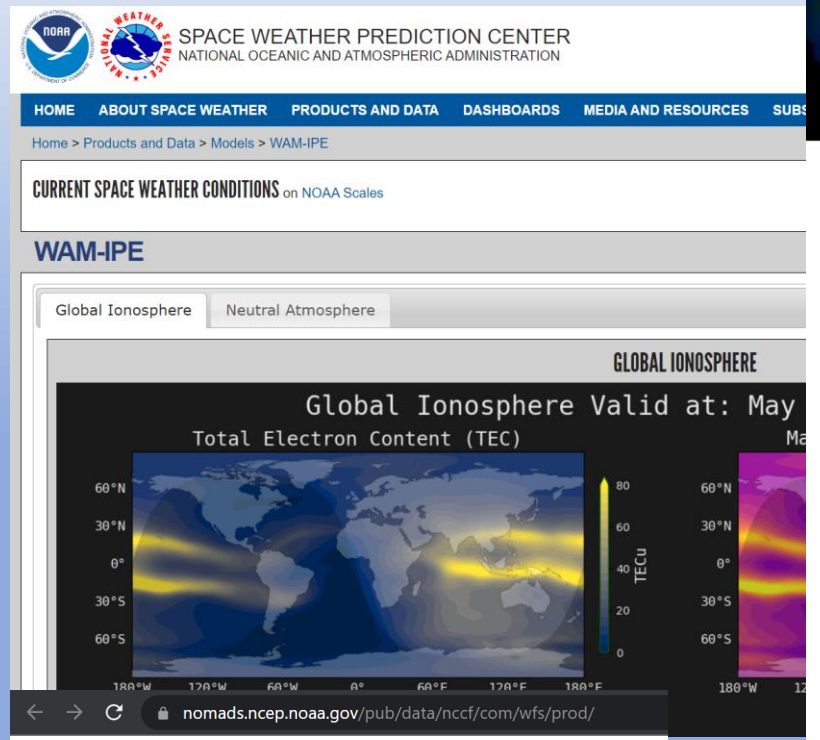
Observational Data

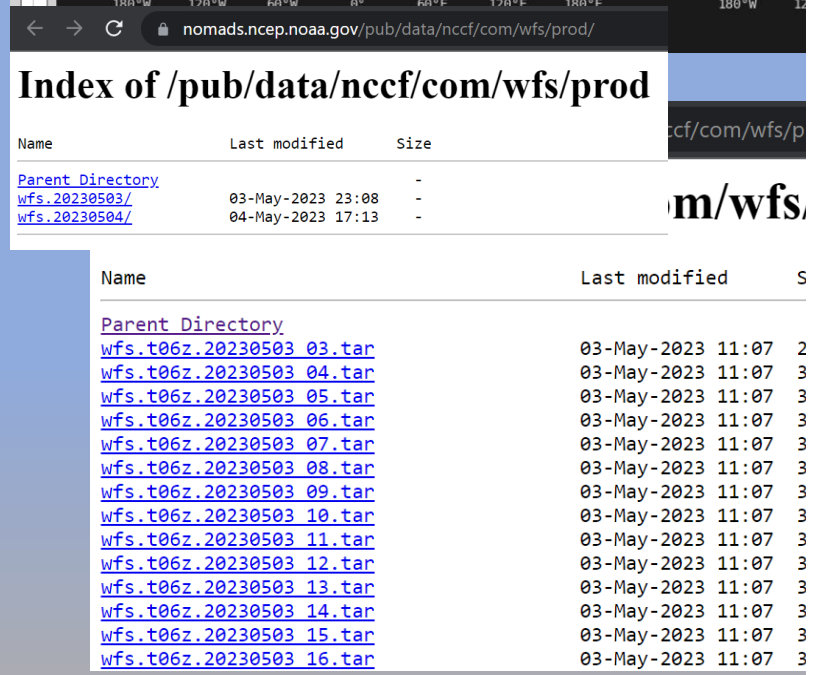
- A growing number of datasets…
  - Are **searchable** through a modern interface (using SPASE),
  - Have **citable** DOIs independent of publications,
  - Are **downloadable** both through web pages and APIs,
  - Are **browsable** via quick-look plots, and
  - Are available on the **cloud**.
- Many observational datasets still lack these properties, so **continued progress is needed**.

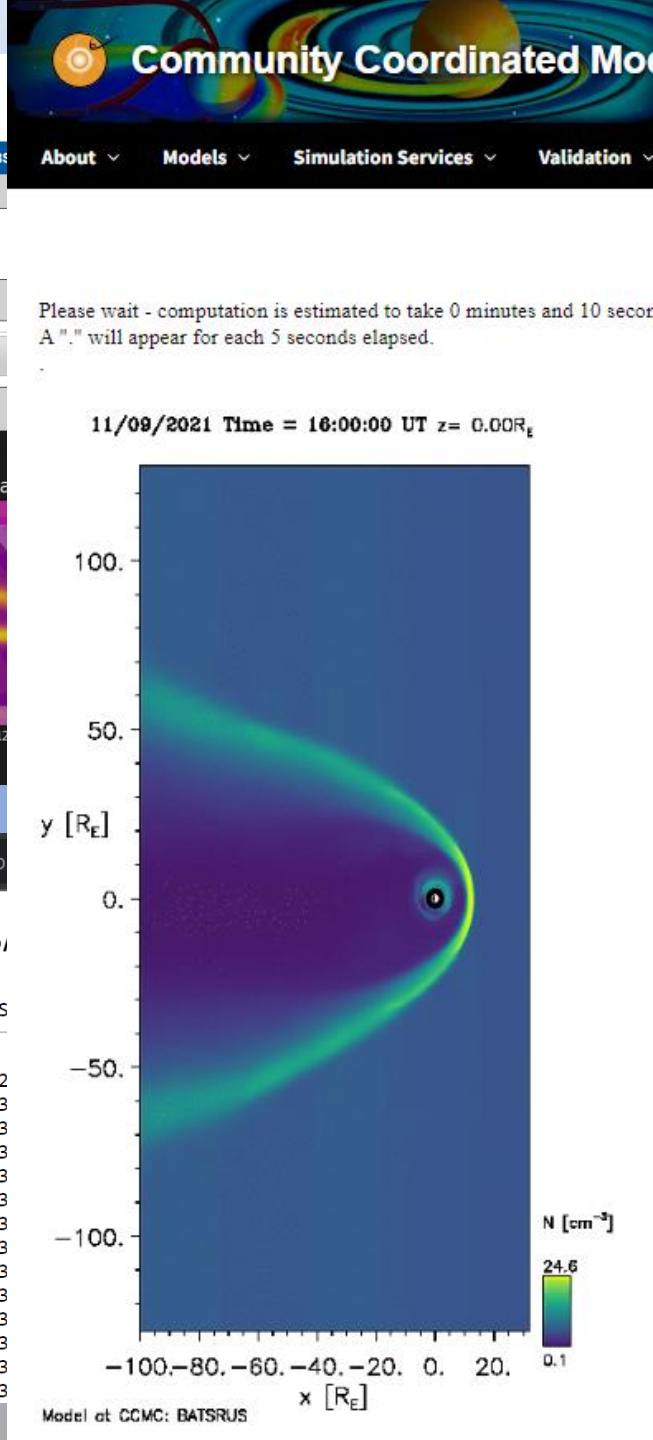


Modeled Data

- 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.





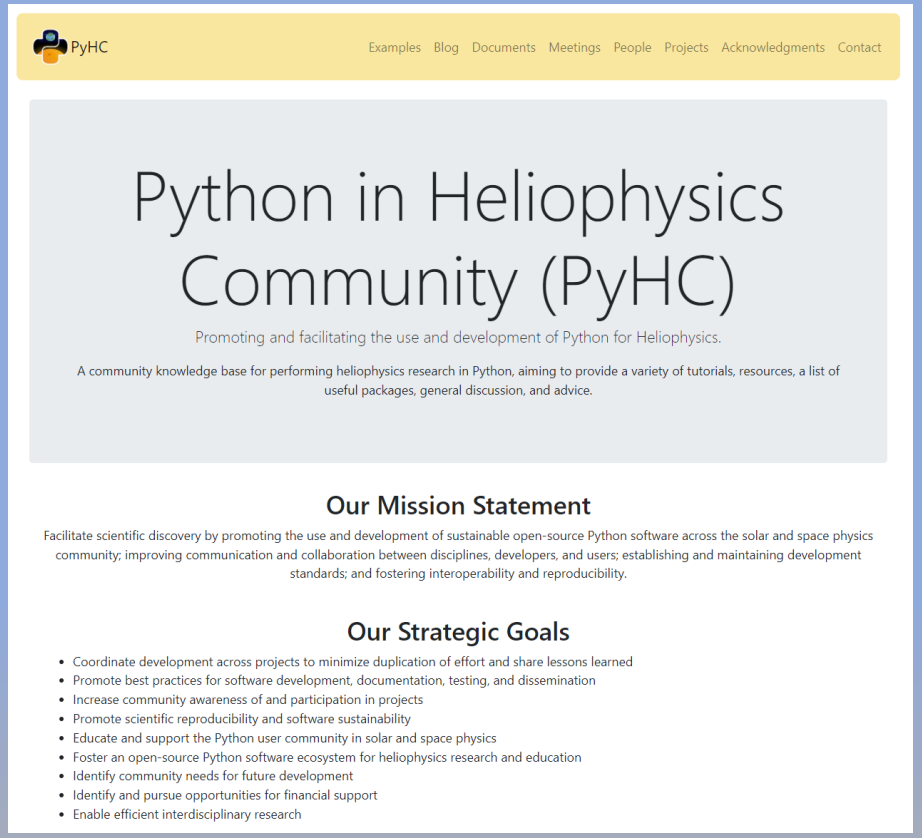


Modeled Data

- 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.
  - **Reproducibility**: 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.
  - **Selection**: Must also include start and stop times, variables contained, coordinate systems, coordinate ranges, domains, and more.

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/



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.