

Data Science and Analytics: Data/Model Resources and Curation Panel

Key Questions:

- 1) Is a more coordinated/sustained data curation effort needed to support R2O2R?
- 2) What data sources are needed but unavailable (proprietary, classified, etc) that are hampering next steps? Do we work to get them available or can (ML, data curation, etc.) take care of it and how?

Moderator: Anthea Coster, Committee

Carrie Black

NSF

William Schreiner

UCAR

Jack Ireland

NASA

Rob Redmon

NOAA

Masha Kuznetsova

NASA/CCMC

Alec Engell

NextGen Federal Systems

Space Weather Operations and Research Infrastructure Workshop: Phase II, Wednesday, April 13, 2022, 1145 ET



NSF Perspective on Data/model resources and curation

Carrie Black

NSF Division of Astronomical Sciences

Program Officer for The National Solar Observatory



Community's Unmet Data Infrastructure Needs

- Easy access. User-friendly;
- Documentation of data for record-keeping and end-users;
- Develop a formal data policy for data citation and attribution;
- File and data standardization. Use a standard/universal format so that data can be readable by one readme file;
- Provide data access reports to data providers and funding agencies for recordkeeping;
- Data repositories. A single site or multiple sites.

Current NSF-Supported Solar and Space Physics Data Systems Infrastructure



- **AST:** Supports data systems at each of its major facilities
- **Madrigal Database (Millstone Hill):** manages and serves archival and real-time data from a wide range of upper atmospheric science instruments.
- **NCAR CISL and HAO:** CISL provides supercomputing, analysis and visualization resources, stores, develops, and curates data sets and maintains user-centered online access. HAO provides data (MLSO and PFI) and models.
- **SuperMAG (APL):** provides easy access to validated ground magnetic field perturbations in the same coordinate system, identical time resolution and with a common baseline removal approach
- **Community Coordinated Modeling Center (with NASA):** provides access to modern space research models; tests and evaluates models; supports Space Weather forecasters; supports space science education

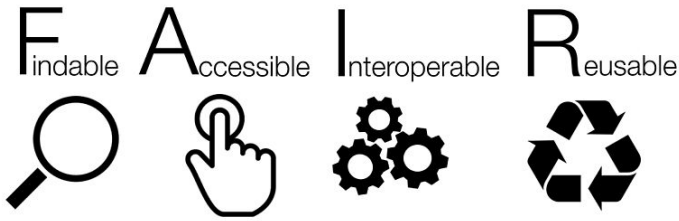


Current Support and Resources

- **FAIROS RCN (NSF 22-553):** Findable Accessible Interoperable Reusable Open Science Research Coordination Networks.
- **Dear Colleague Letter: Effective Practices for Making Research Data Discoverable and Citable (Data Sharing) (NSF 22-055):** describes and encourages effective practices for publicly sharing research data, including the use of persistent digital identifiers (PDIs).
- **EarthCube (NSF 21-515):** supports development of a Community-Driven Data and Knowledge Environment for the Geosciences
- **Geoinformatics (GI) (NSF 21-583):** supports development and implementation of sustainable funding models to preserve data and software products of value to Earth Science research.
- **Dear Colleague Letter: Pilot Projects for Cyberinfrastructure Centers of Excellence (NSF 21-037):** CI CoEs are service-oriented hubs of expertise and innovation targeting specific areas, aspects, or stakeholder communities.



NSF Data Policies & Data Management Plan



- All core and targeted programs in the NSF follow the data policies in the Proposal and Awards Policies & Procedures Guide (PAPPG);
- Each division also has their own additional policies:
 - AGS - <https://www.nsf.gov/geo/geo-data-policies/ags/index.jsp>
 - AST - <https://www.nsf.gov/bfa/dias/policy/dmpdocs/ast.pdf>

Current Challenges

- Review of the DMP is not uniform across NSF or even within programs
- The quality of the DMP frequently does not play a significant role in the merit review and proposal funding recommendation;
- Report on the DMP performance is not required or enforced in the project reports;
- GS has encouraged but not required the DMP to be in alignment with the FAIR principles.



2024 Solar and Space Physics Decadal Survey

3. **Develop a comprehensive ranked research strategy that provides an ambitious, but realistic, approach to address these science goals.** The strategy will include consideration of:
 - a. The combination of ground- and space-based investigations to enhance progress on the prioritized science goals.
 - b. Data and computing infrastructure needed to support the research strategy and the long-term utility, usability, and accessibility of acquired data;**
 - c. Technical, risk, and cost assessments of recommended major investments, when deemed useful;
 - d. Decision rules that can accommodate reasonable projected budget deviations or changes in activities' urgency; and
 - e. The international landscape, inter-agency collaborations, public-private relationships, and innovative partnerships.



Actionable Deliverables

The Community can consider

- Participating in the decadal survey
- Developing training courses on how to use data resources to make them more accessible and open access.
- Forming working groups to work on specific topics: data standardization, data and documentation formats, and data policy for data citation and attribution.

NSF side is

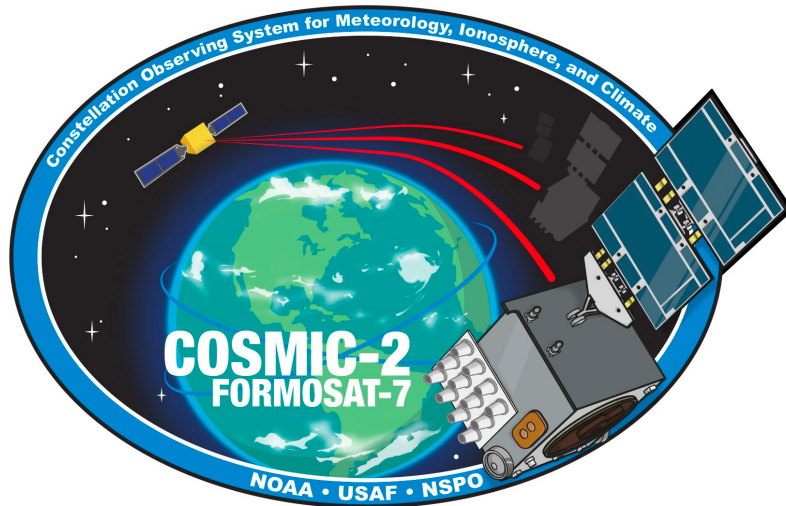
- collaborating with other federal agencies to leverage resources and coordinate requirements.
- open to sponsoring community workshops on Data Systems Infrastructure
- working to develop a data policy more in line with the FAIR principles.
- Investigating how to improve PI reporting DMP practices in the project reports.

National Academy of Science Engineering and Medicine

Space Weather Workshop Phase II

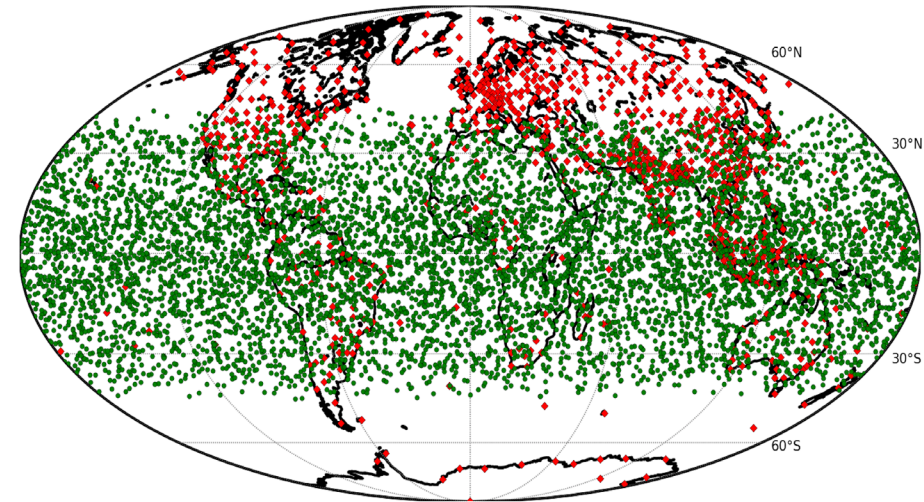
Panel: Data Science and Analytics

Data/Model Resources and Curation



Bill Schreiner (schrein@ucar.edu)
NOAA COSMC-2 Mission Scientist
IROWG Space Weather Subgroup
UCAR COSMIC Program
Project Scientist

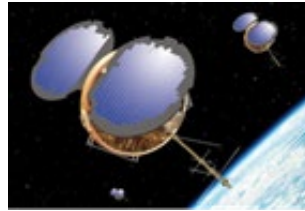
April 13, 2022



GNSS Radio Occultation From Research to Operations



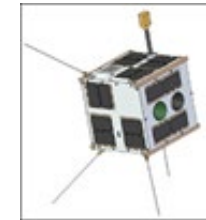
GPS/MET - 1995: NSF -
Funded,
Partners: NASA/JPL, U of
Arizona, Cal Tech, Orbital
Sci., NCAR



COSMIC - 2006: Funded by NSPO (Taiwan),
NSF, NOAA, NASA, Air Force, Navy
Partners: JPL, Orbital Sci.



COSMIC-2 - 2019: Funded by NOAA,
Air Force, NSPO (Taiwan),
Partners: NASA/JPL, UT-Dallas



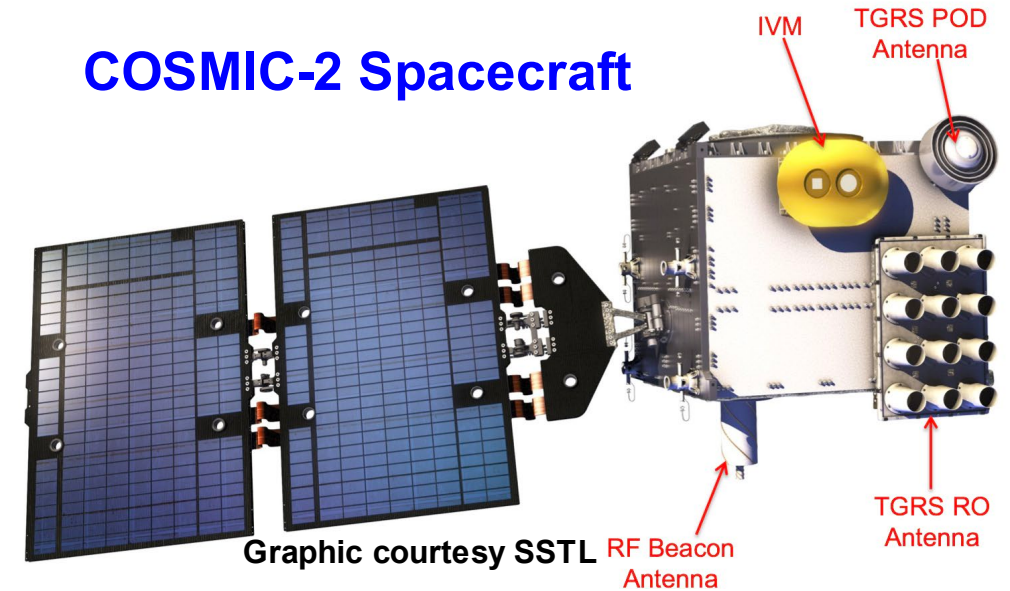
Commercial RO –
2020 and beyond



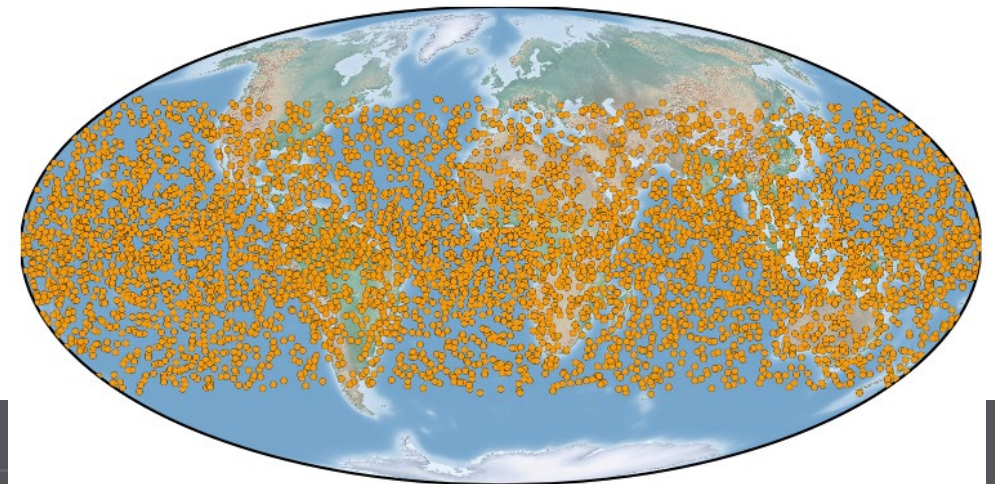
GNSS RO Overview

- Global Navigation Satellite System (GNSS) radio occultation (RO) is now a critical component of the global observing system - enabling significant improvements for weather forecasting, space weather, and climate applications
- NOAA's six-satellite COSMIC-2 GNSS RO sounding mission
 - Launched June 2019
 - COSMIC-2 is now collecting ~6,000 of the highest quality soundings/day with ~27 min latency
 - Providing uniform geographic coverage for low/mid latitudes with better than 6 hr revisit times
- Commercial RO Data Purchases
 - NOAA DO1 (2020), DO2 (2021), and DO3 (2022), (DO3 to provide ~6,000 soundings/day global)
 - NASA, 2020, 2021
 - DoD

COSMIC-2 Spacecraft



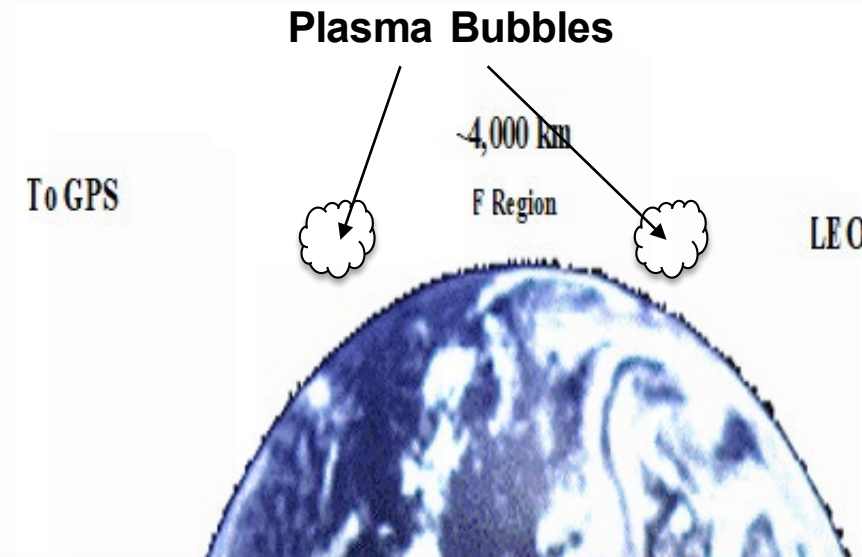
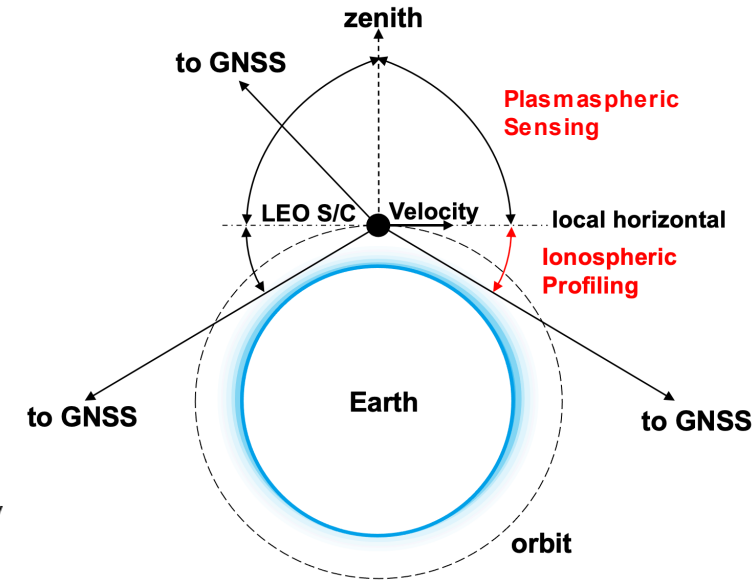
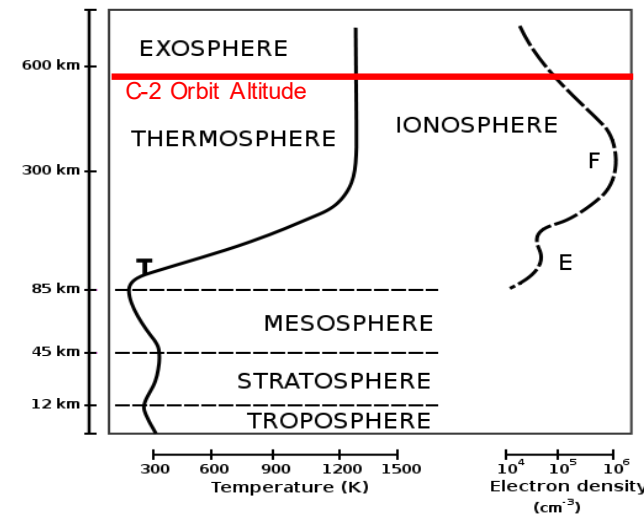
COSMIC-2 on Feb 5, 2020 – 5,546 profiles



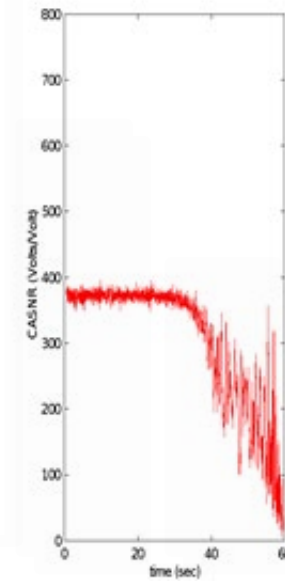
GNSS RO Space Weather Observations



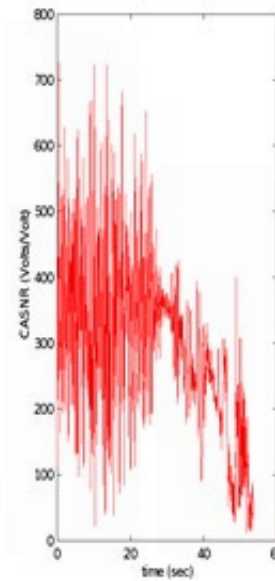
- Large- and small-scale variations of the ionospheric electron density are important to monitor and predict, because they affect radio-wave propagation and can impair satellite communication, navigation and electric power grids
- **GNSS RO Profiling**
 - The COSMIC-2 spacecraft track GNSS satellite signals at negative elevation angles as they descend through the ionosphere, i.e., measure Total Electron Content as a function of altitude with accuracy of $< 3 \times 10^{16}$ electrons/m² (Pedatella et al., 2021)
 - The TEC as a function of altitude is then used to compute a profile of electron density (el/cm³) with accuracy of ~6-7% (Cherniak, et al., 2020)
- **GNSS RO Scintillation**
 - The COSMIC-2 RO instruments measure scintillation indices for every tracked GNSS satellite in view, and high-rate data that can be used to accurately geolocate plasma bubbles with ~100 km uncertainty
 - New plasma bubble map and All Clear products are now being developed



No scintillation



Scintillation



Questions Provided by Organizers



1. *Is a more coordinated/sustained data curation effort needed to support R2O2R? YES*
- Continue developing standard data and metadata formats for RO:
 - COSMIC-2 data (netCdf with Climate/Forecast Conventions) are currently being archived at UCAR and NCEI
 - Develop standard data formats for TEC/Scintillation space weather data (WMO BUFR, other?)
 - Develop metadata standards that includes: payload state-of-health (SOH), configuration and software versions, spacecraft SOH and performance, payload tracking/retrieval algorithm documentation, data processing software documentation/versions, etc (WIGOS metadata, other?)
 - Work with community to develop AI/ML-ready RO data and metadata formats
- Large data volumes and models require development of data proximate cloud computing environments for RO data processing and science applications
- Consider development of community-developed space weather assimilative models



Questions Provided by Organizers



2. What data sources are needed but unavailable (proprietary, classified, etc) that are hampering next steps? Do we work to get them available or can machine learning, data curation, etc. take care of it, and how?

- Continuity for GNSS RO
 - NOAA backbone constellation with additional commercially-purchased data?
 - Backbone: high-quality, reliable, long-lasting, well-documented
- GNSS RO profiling datasets
 - Future RO missions should also track negative elevation angles to enable ionospheric profiling
- Low latency datasets
 - Future missions should provide low latency data (approach real-time)
 - IROWG Space Weather Subgroup and CGMS Space Weather Coordination Group are meeting in early May to discuss requirements/provision of low latency data
- Pursue commercial RO Level-0 data, metadata, formats, readers, documentation
- Work with science community to develop useful gridded products



References/Acronyms



References

Cherniak, et al., (2020) Accuracy Assessment of the Quiet-time Ionospheric F2 peak Parameters as Derived from COSMIC-2 multi-GNSS Radio Occultation Measurements, J. Space Weather Space Clim, doi: <https://doi.org/10.1051/swsc/2020080>

Pedatella, et al., (2021) Processing and Validation of FORMOSAT-7/COSMIC-2 GPS Total Electron Content Observations, Radio Science, <https://doi.org/10.1029/2021RS007267>

Acronyms

- COSMIC = Constellation Observing System for Meteorology, Ionosphere, and Climate
- GNSS = Global Navigation Satellite System
- IROWG = International Radio Occultation Working Group
- IVM = Ion Velocity Meter
- NSPO = National Space Organization
- RO = Radio Occultation
- TEC = Total Electron Content



NASA Heliophysics Science Data

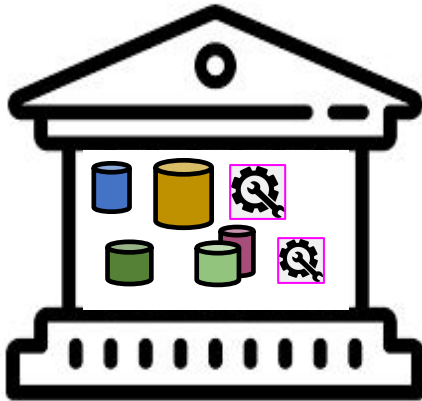
Jack Ireland

Project Scientist, Solar Data Analysis Center (SDAC)
US Project Scientist, SOHO

Heliophysics Digital Resource Library (HDRL) : High Level Strategy

“Preserve”

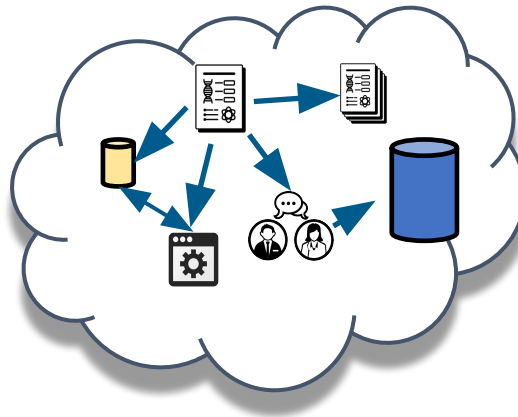
Enhance Archives and Services



Maintain and improve existing archives and services in light of increasing demands

“Discover”

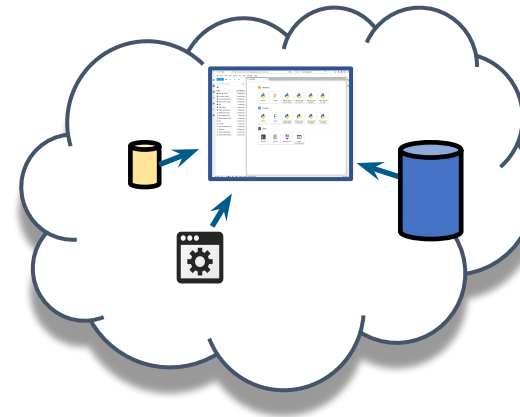
Enhance Discoverability



Increased interlinking of research artifacts, ADS integration, DOIs, improved standards, etc

“Explore Further”

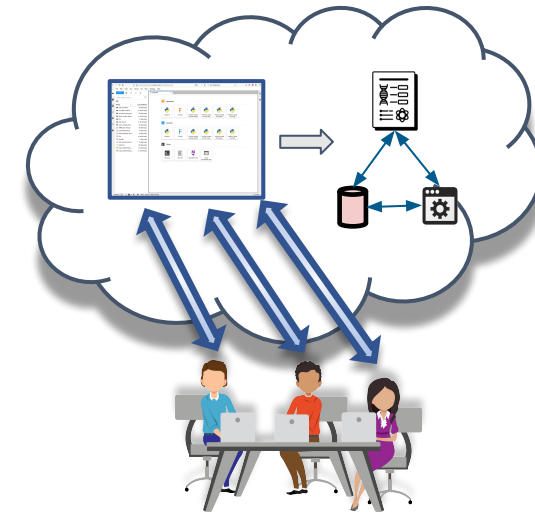
Unlock Big Data Research



High End Compute close to large (up to 100 Tb) and Big Data (~Pb) with software support (ML, PyHC, etc). Leverage earth science platform and expertise.

“Share & Publish”

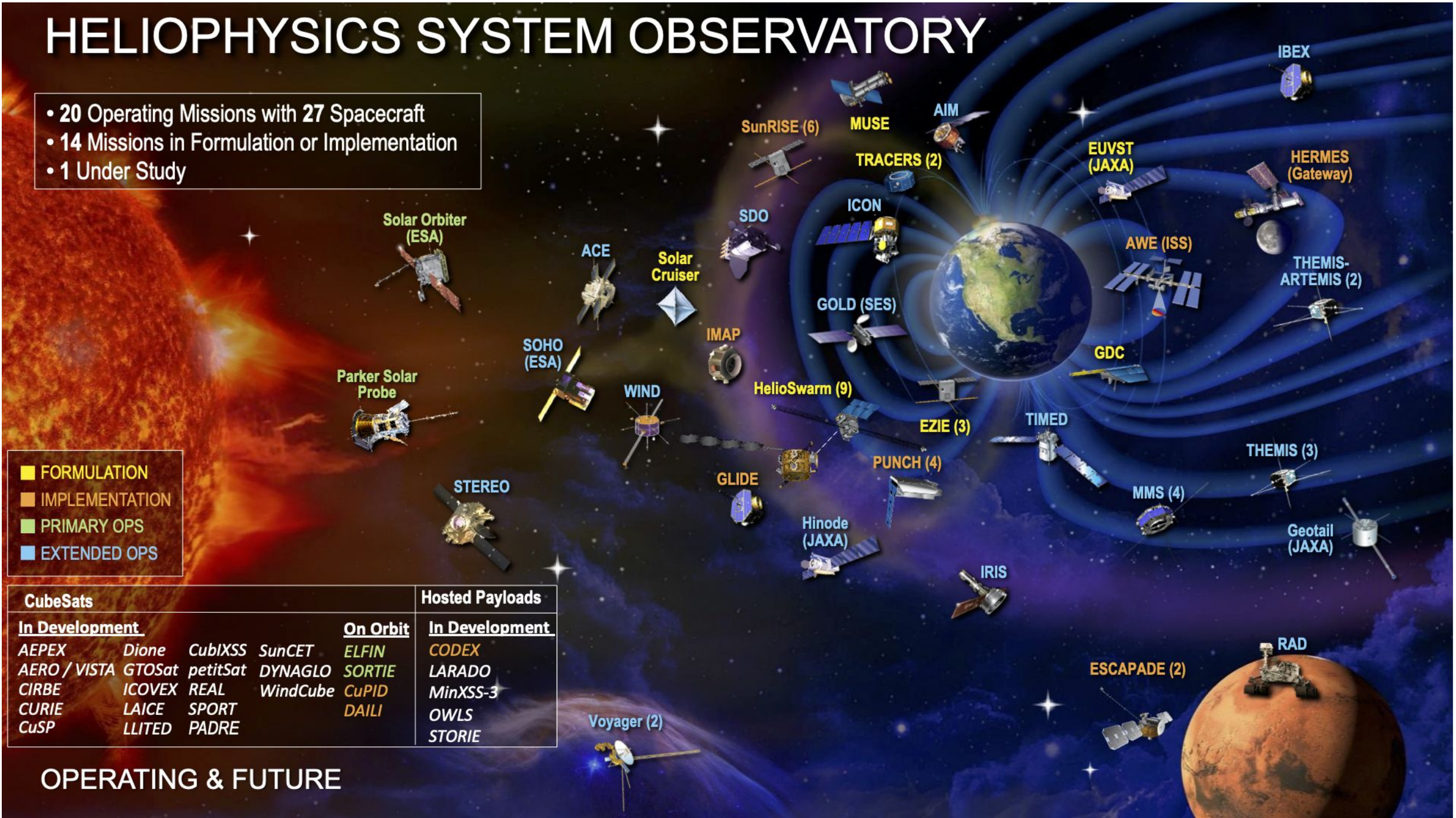
Enable Team Open Science



Collaborative Online Research, Compute, and Publishing Platform

HELIOPHYSICS SYSTEM OBSERVATORY

- 20 Operating Missions with 27 Spacecraft
- 14 Missions in Formulation or Implementation
- 1 Under Study



- FORMULATION
- IMPLEMENTATION
- PRIMARY OPS
- EXTENDED OPS

CubeSats

In Development

AEPEX Dione CubIXSS SunCET
 AERO / VISTA GTOSat petitSat DYNAGLO
 CIRBE ICOVEX REAL WindCube
 CURIE LAICE SPORT
 CuSP LLITED PADRE

On Orbit

ELFIN
 SORTIE
 CuPID
 DAILI

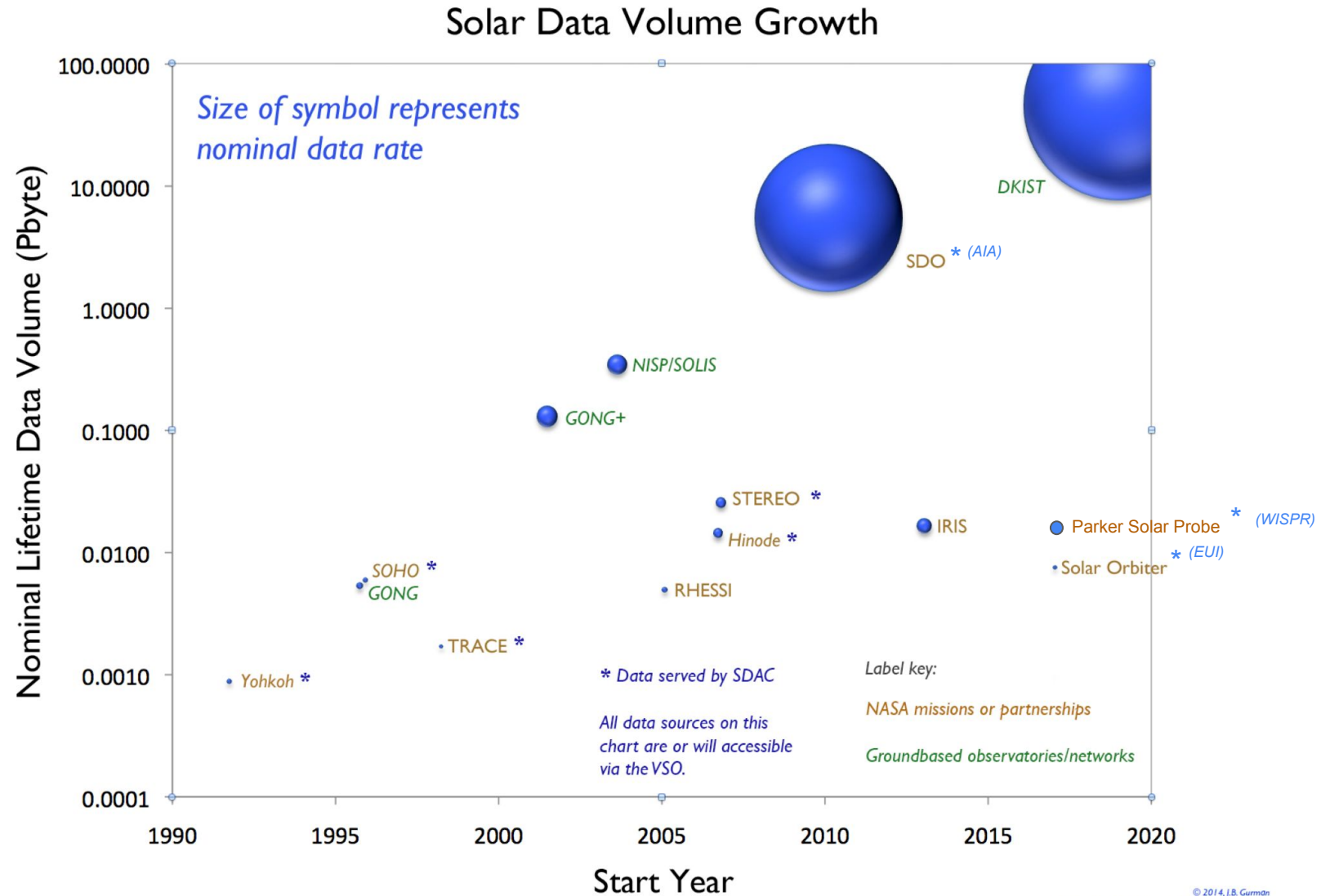
Hosted Payloads

In Development

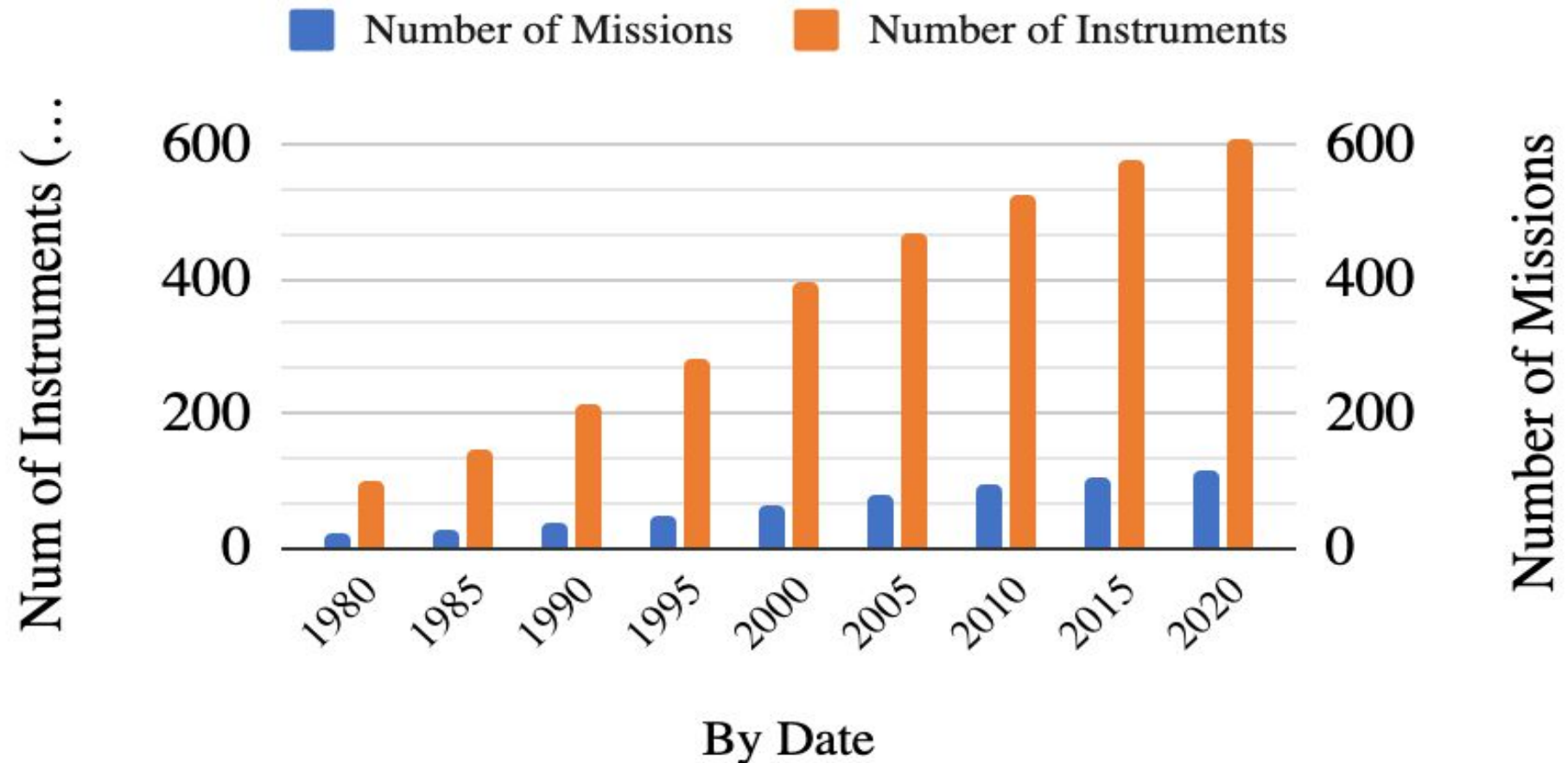
CODEX
 LARADO
 MinXSS-3
 OWLS
 STORIE

OPERATING & FUTURE

Solar physics science data - volume



Space physics science data - variety



Total volume ~ 0.4PB

Directions for science data infrastructure

- NASA led meeting of NASA cross-divisional data infrastructure representatives (August 2018)
 - Only meeting (so far) of representatives from all science divisions who design/implement science data infrastructure.
 - Issues of volume and variety in heliophysics exist in other divisions.
- NASA HQ led exercise involving SDAC, SPDF, CCMC and HDMC (2019)
 - Series of recommendations/ideas generated through a series of meetings at NASA HQ and GSFC
- Community involvement via a user workshop (2021)
 - Invited members of the heliophysics community to share their thoughts/experiences when interacting with NASA heliophysics data infrastructure.
 - Themes identified
- NASA RFI (ending February 2022)
 - Canvassed the entire heliophysics community
 - Inputs being worked on at NASA HQ

Heliophysics Digital Resource Library (HDRL)

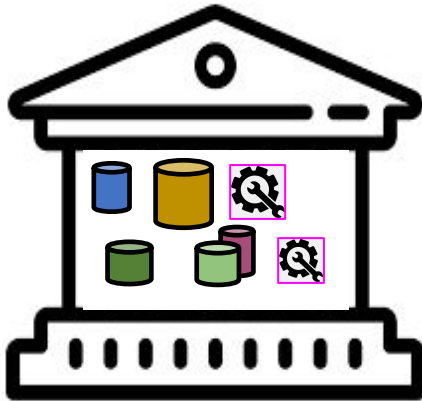
Vision : User-Driven Acceleration of Heliophysics Research

- Enabling Open Science
- Lower Current Barriers to Doing Research
- Implement New Critical Capabilities
- Responsive to Changing Community Needs

HDRL : High Level Strategy

“Preserve”

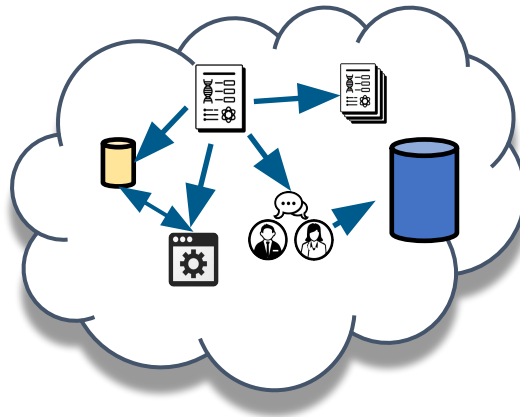
Enhance Archives and Services



Maintain and improve existing archives and services in light of increasing demands

“Discover”

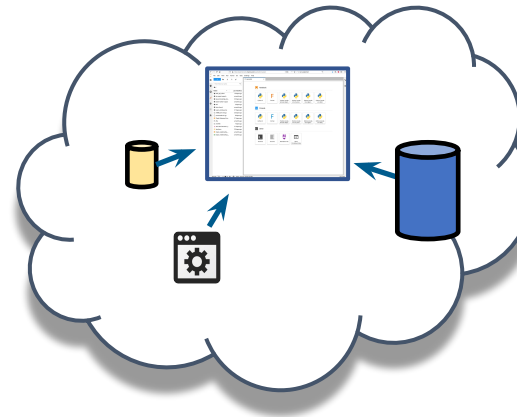
Enhance Discoverability



Increased interlinking of research artifacts, ADS integration, DOIs, improved standards, etc

“Explore Further”

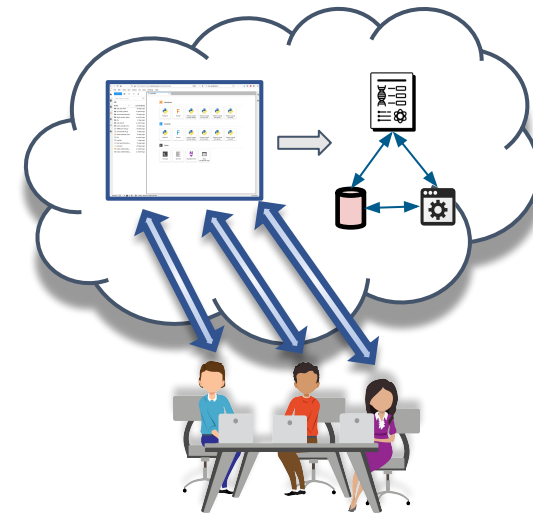
Unlock Big Data Research



High End Compute close to large (up to 100 Tb) and Big Data (~Pb) with software support (ML, PyHC, etc). Leverage earth science platform and expertise.

“Share & Publish”

Enable Team Open Science



Collaborative Online Research, Compute, and Publishing Platform

HDRL : science data infrastructure components

HP Data and Model Consortium

D. Aaron Roberts (PS), Brian Thomas (DPS)

Registries and DOIs for all digital resources; SPASE Data Model.

Heliophysics Data Portal (HDP; including solar)

Python and other software integration (PyHC). Analysis and visualization services ((Py)SPEDAS, Autoplot).

Data upgrades and services.

HelioCloud initiative with data and software from all groups.

Solar Data Analysis Center (SDAC)

Jack Ireland (PS)

Solar Data Final Active Archive for NASA (and other) solar missions.

Virtual Solar Observatory data access.

Helioviewer. SunPy. SolarSoft.

High Performance Computing for NASA HP.

Space Physics Data Facility (SPDF)

Robert Candey (PS), Lan Jian (DPS)

Non-solar Data Final Active Archive for NASA (and other) missions.

CDAWeb data browsing and access; Web Service access.

OMNIWeb data production and serving.

SSCWeb and 4-D spacecraft orbit facility. Common Data Format.

Collaborators

Community Coordinated Modeling Center Center for HelioAnalytics (GSFC)

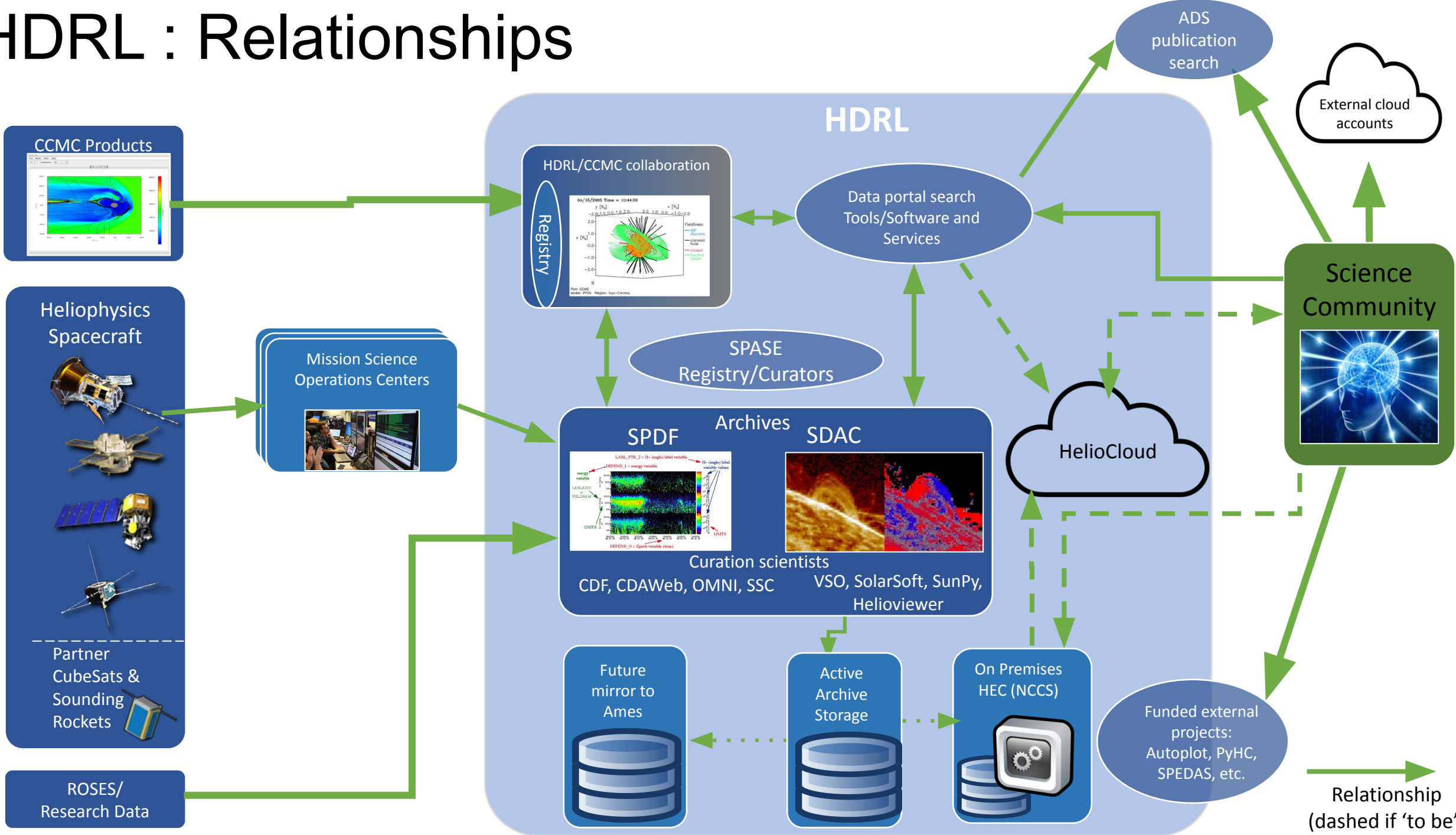
Data—model comparisons.

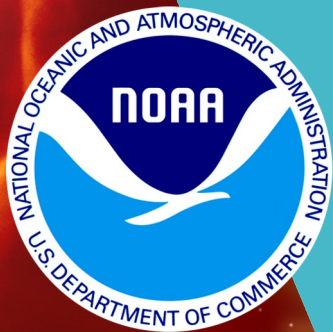
Registry of models and output.

“Kamodo” enabled visualization.

AI/ML cloud computing

HDRL : Relationships

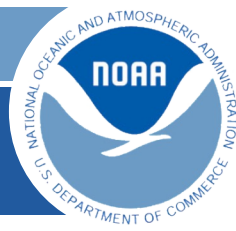




NOAA
National Satellite and
Information Service

Panel Data Science and Analytics: Data/model resources and curation

Rob Redmon, NOAA NCEI and NCAI
National Centers for Environmental Information
NOAA Center for Artificial Intelligence
April 13, 2022



Questions Provided by Organizers

1. Is a more coordinated/sustained data curation effort needed to support R2O2R?

- Data/model curation and standards are living processes so there's always more we can do to coordinate efficiently, e.g. participate in standards development.
- Suggest using new programs such as NASA's [HDRL](#), and NOAA's Space Weather Follow On (SWFO) to lean into interoperable data/metadata standards (e.g. SPASE) and existing interfaces (e.g. [HAPI](#), [DSCOVER](#)), and developing benchmark datasets akin to augmented Climate Data Records.
- Leverage USG/Academia/Industry collaborations, e.g. Earth Science Information Partners (ESIP) [Data Readiness](#) and [ML](#), and others.
- Resources: [SWx R2O and O2R Framework](#), [NCEI/STP](#), [NOAA Archive Guide & Process](#)



Questions Provided by Organizers

2. What data sources are needed but unavailable (proprietary, classified, etc) that are hampering next steps? Do we work to get them available or can machine learning, data curation, etc. take care of it, and how?

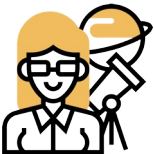
- Retirement of DMSP and POES/Metop SWx imagers, particle & field sensors.
- Data Science can teach us much (e.g. feature/anomaly detection, model emulation), but can't replace observations in a data sparse environment.
- Continue to work together to bring currently unavailable and hard to use data efficiently to the research community to drive future R2O2X, e.g. via the Space Weather Operations, Research and Mitigation Subcommittee (SWORM).
- NOAA is working to migrate all of it's data holdings to the Cloud, embracing Findable Accessible Interoperable Reusable (FAIR), Big Data Program.
- Does AL/ML have a role anonymizing source data sufficiently for proprietary, or protected concerns, e.g. perhaps via OADR Proving Ground for Space Situational Awareness, testbed hack-a-thons?
- Resources: [SWORM NSW-SAP](#), [APL NASA Gap Analysis](#)



NOAA NCEI Space Weather

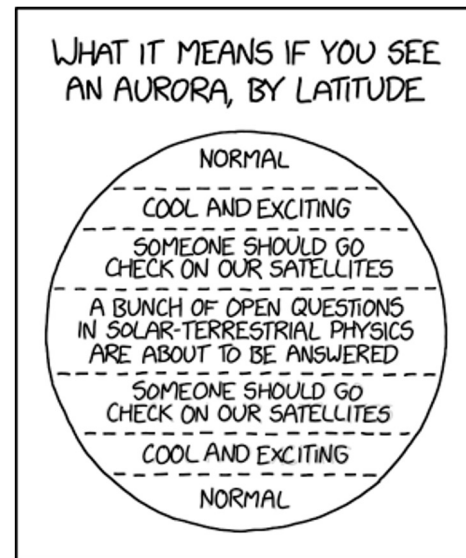
Serving the Research Community and Operational Needs

NCEI serves Retrospective & Science Users



- Science Research - Academic, NOAA, DOD, NASA, NSF, ...
- Ops Research - SWPC, NASA, DOD, other Ops users, Industry
- Education & Outreach - Students & Educators
- Program tailored Operational development support in support of NOAA/NESDIS new Space Weather Next Program, e.g. Algorithm/Product development, Calibration & Validation
- Data access, archive & scientific authoritative stewardship
- Scientific research (R2O/O2R)

xkcd comic [#2233](#)



Adapted from Fall AGU 2021 NCEI Town Hall.

NWS Space Weather Prediction Center serves Ops Users



- NOAA/NWS - Space Weather Prediction Center, International Space Environment Service of forecasters, *and their customers around the world*: airlines, power grid, DOD, GPS systems, satellite providers, etc



NCEI Stewardship & Scientific Products Future

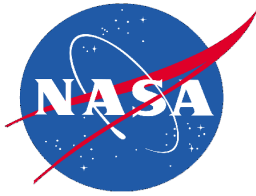
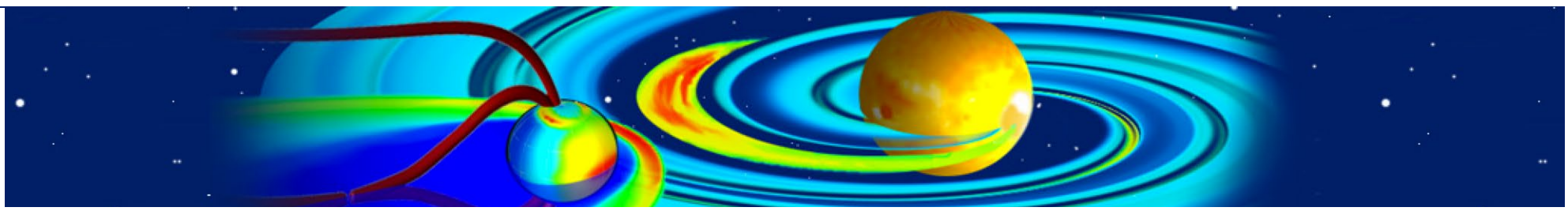
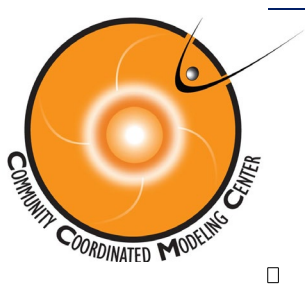
Overall:

- Recent survey from ESIP on needs to achieve AI-readiness for open environmental data → half of respondents spend **at least 50% of their time data massaging**, before they can begin. See [AI-ready Checklist](#).
- X-Prizes: In partnership with the NASA Tournament Lab, NOAA conducted a real-time crowdsourcing magnetic field modeling challenge.
- NOAA has [published strategies](#) and plans for harnessing the best technologies: Cloud, Data, AI (noaa.gov/ai), Citizen Science, etc.

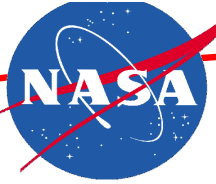
NCEI's SWx Archives:

- NCEI is moving toward the cloud for storage and access of our data holdings. [DSCOVR](#) and [GOES-R](#) are our heritage as we prepare for SWFO. What features would you like to see included in this transition?
- How can NCEI better support you as a data user?
- What SWx products should NOAA work to include in our future portfolio?





Community Coordinated Modeling Center



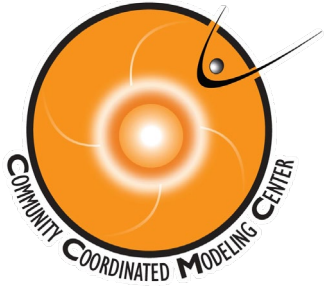
Access point to state-of-the-art space environment modeling capabilities

Portal to Research-to-Operations (R2O) transition pipeline

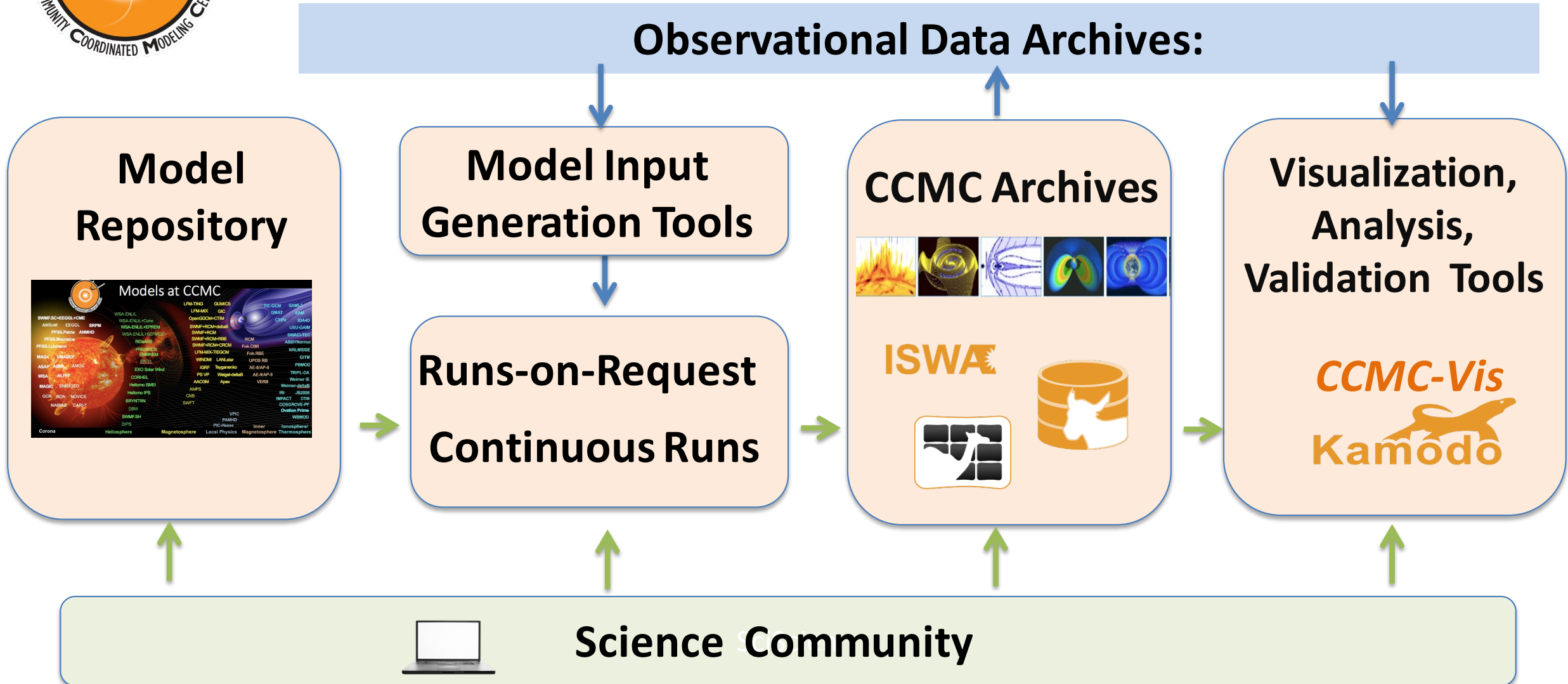


M. Kuznetsova & CCMC Team

MODELS • DATA • TOOLS • DATABASES • FRAMEWORKS • SERVICES



CCMC Models, Tools & Services at a Glance





CCMC-SWPC R2O2R Pipeline:

CCMC space weather information (as is)



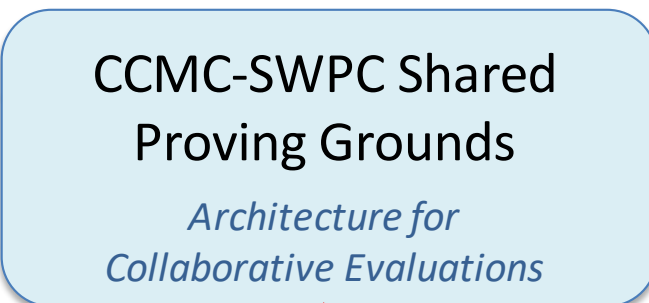
O2R: Enabling community access to archives of operational data streams, model inputs and simulation outputs through CCMC systems

*Research,
educational users
of space weather
information (as is)*



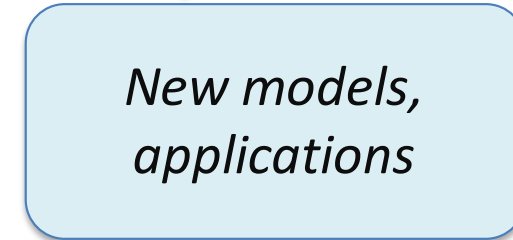
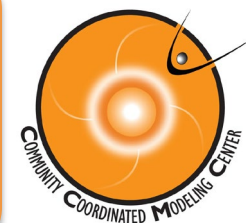
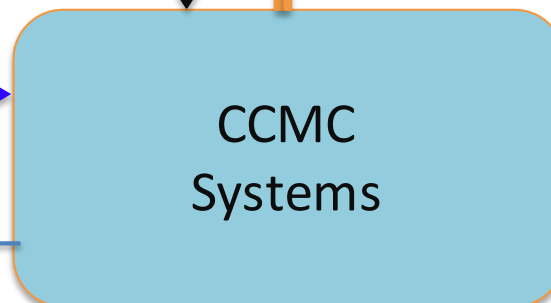
*shared
repository*

*operational
models*



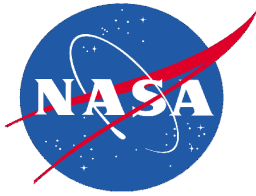
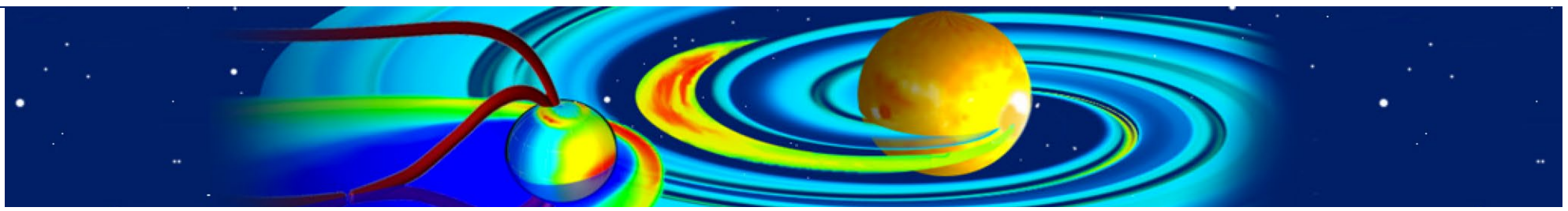
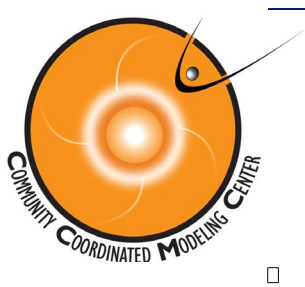
*shared
repository*

*R2O
transition*



Vision for Modernization

- Establish and follow **best practices** for on-boarding and implementation
 - **Shared environments** on the AWS cloud and at NASA HPC systems.
 - Increase modeler participation. Move towards plug and play
 - Utilizing **shared GitLab** repositories
 - Utilize **container** technology to improve portability.
- Improve **quality** of simulation archives
 - Follow **SPASE metadata** standards for all information in CCMC archives
 - Implement **HAPI** standards and **APIs**
 - Include information about model capabilities into metadata
- Enable **reproducibility** of simulation runs and plots
- Improve **robustness** and **speed** of simulations (GPU)
- Initiate **open-source** pilot projects (e.g., plug and play with code portions)
- Address journals **simulation outputs accessibility** requirements: store results of simulations performed outside of CCMC and utilize CCMC tools for visualization and analysis.



Community Coordinated Modeling Center



Access point to state-of-the-art space environment modeling capabilities

Portal to Research-to-Operations (R2O) transition pipeline



M. Kuznetsova & CCMC Team

MODELS • DATA • TOOLS • DATABASES • FRAMEWORKS • SERVICES



CCMC-SWPC R2O2R Pipeline:

CCMC space weather information (as is)



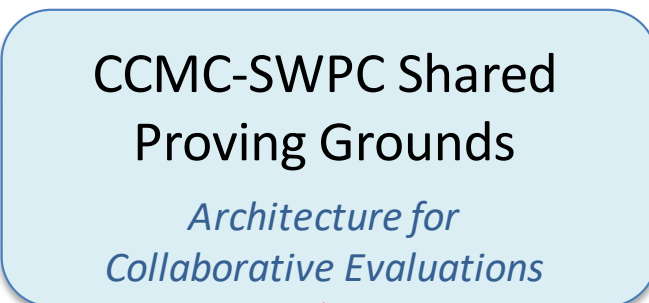
O2R: Enabling community access to archives of operational data streams, model inputs and simulation outputs through CCMC systems

*Research,
educational users
of space weather
information (as is)*



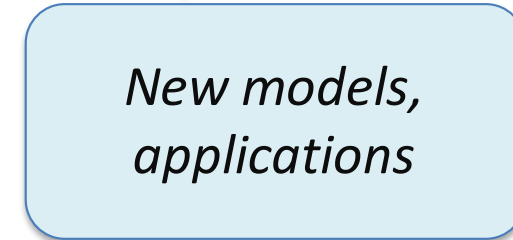
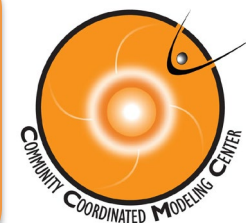
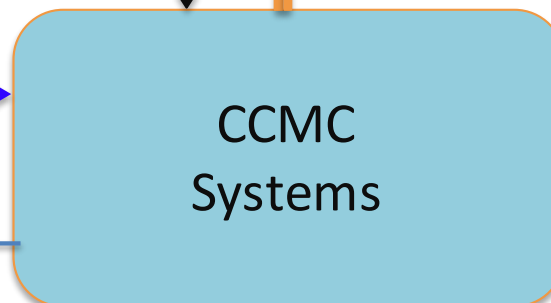
*shared
repository*

*operational
models*



*shared
repository*

*R2O
transition*



Vision for Modernization

- Establish and follow **best practices** for on-boarding and implementation
 - **Shared environments** on the AWS cloud and at NASA HPC systems.
 - Increase modeler participation. Move towards plug and play
 - Utilizing **shared GitLab** repositories
 - Utilize **container** technology to improve portability.
- Improve **quality** of simulation archives
 - Follow **SPASE metadata** standards for all information in CCMC archives
 - Implement **HAPI** standards and **APIs**
 - Include information about model capabilities into metadata
- Enable **reproducibility** of simulation runs and plots
- Improve **robustness** and **speed** of simulations (GPU)
- Initiate **open-source** pilot projects (e.g., plug and play with code portions)
- Address journals **simulation outputs accessibility** requirements: store results of simulations performed outside of CCMC and utilize CCMC tools for visualization and analysis.



NEXTGEN
FEDERAL SYSTEMS

Space Weather II Workshop

Panel: Data Science and Analytics Data/Model Resources and Curation

National Academy of Science Engineering and Medicine

April 13, 2022

Alec Engell (aengell@nextgenfed.com)

Principal Data Scientist



Observations of Data/Models Resources and Curation

Is a more coordinated/sustained data curation effort needed to support R202R?: **Yes**

How? (high level):

Infrastructure

- Software, Formats, Storage, Compute
 - Common interfaces, tools, adaptors

Execution

- Agile, DevSecOps, CI/CD
 - Efficient evolution with changing requirements and technology/methods is key

Roadmaps and data architectures are not always shared across the space weather and heliophysics communities

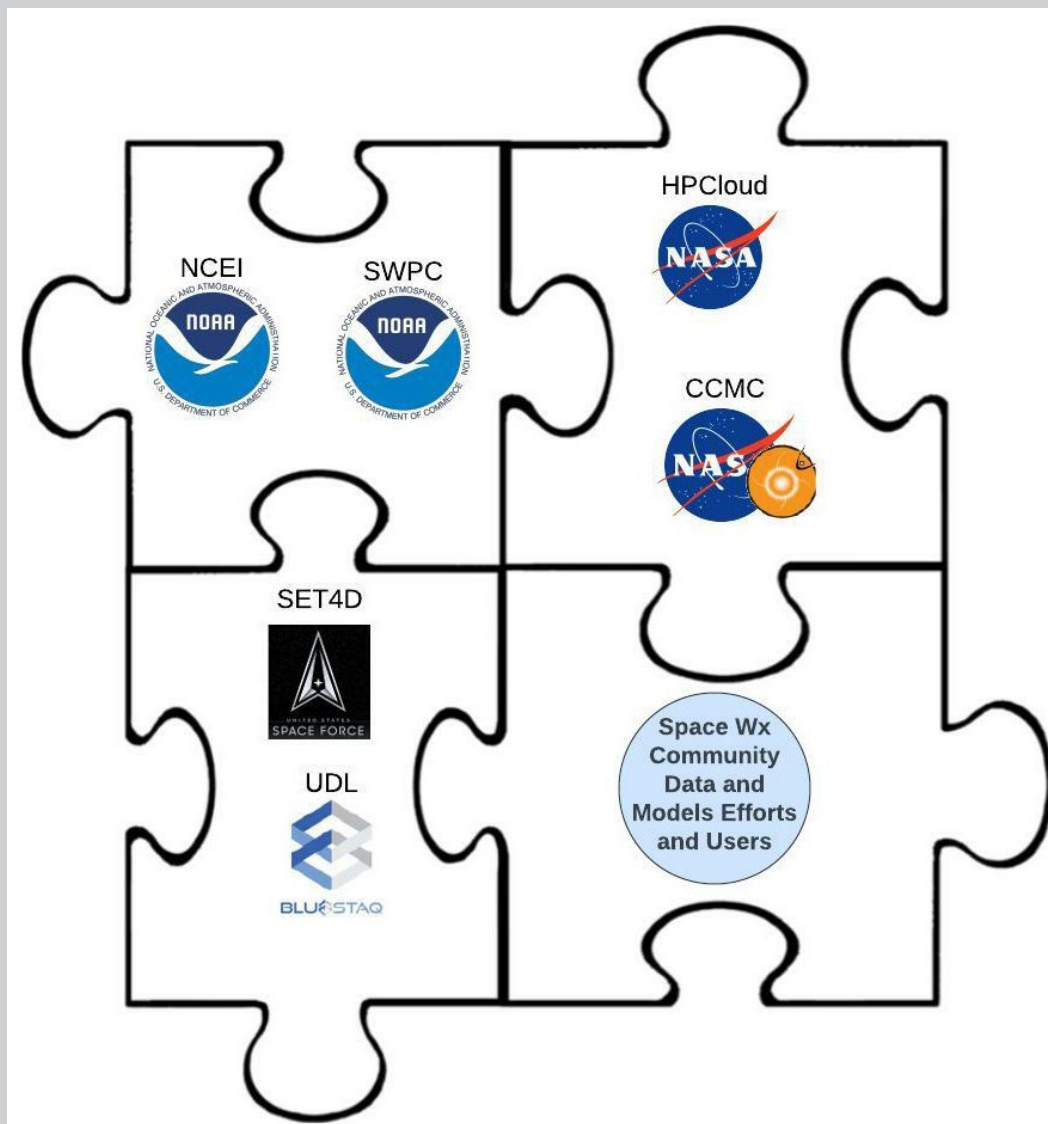
- Reinventing the wheel will occur
- Unlikely to align: Will make R20-O2R difficult; R20 and O2R are often one-way streets
 - Must couple!!
- Potential solution: Create a small focused team of the leads taking on these challenges with outside expertise



Data/Model Resource Curation Efforts

Not limited to:

- NASA CCMC is the “Proving Grounds” for R20 to SWPC testbed
- NASA Heliophysics Digital Resource Library (HDRL) and the HPCloud environment
- NOAA SWPC Testbed and operational model outputs
- NOAA NESDIS/NCEI: archive and access of solar and space environmental data and derived products collected by NOAA observing systems
- Space Environment Toolkit for Defense (SET4D) and the Unified Data Library (UDL)
- Pangeo: a community of people working collaboratively to develop software infrastructure to enable Big Data geoscience research → Panhelio? (pangeo.io)





Next Steps for Unavailable Data Sources

“What data sources are needed but unavailable that are hampering next steps?”

- Roadmaps, architectures and resources are not always known/available/documented
- Operational models and forecast data are not often accessible
 - “How do I know if my capability is improving upon the state-of-the-art ops forecasts?”

“Do we work to get data sources available (proprietary, classified, Big Data)?”: **Yes**

- Research groups can work to get clearances when on appropriate projects (hard!)
- Research groups and businesses can partner with business that have clearances
- Businesses and universities can make their IP available for research purposes only
 - Must try to adhere to R2O-O2R repeatable workflows
- Data virtualization, formats, and supporting software (e.g., Pangeo → Panhelio)

“Or, (**And**) can machine learning, data curation, etc. take care of it?": **Partly**

“How?” (part of the picture!):

- Large datasets can be virtualized in the cloud
- Software can support larger than memory compute
- Provide customizable tools at the location of data data → Analysis, ML, and model-ready data
- Bring the user to the data with necessary tools.



General Data Challenges





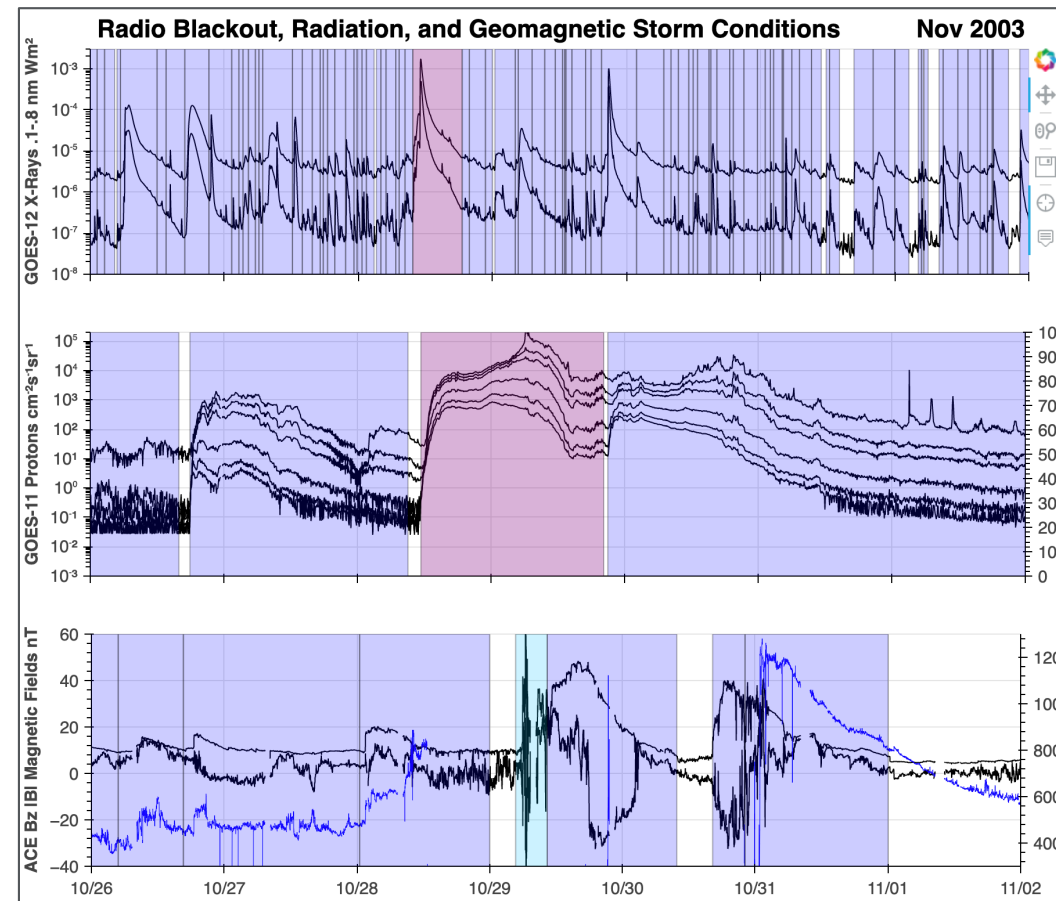
Specific Space Wx Data Challenge Examples

- **Our community**
 - Data spans multiple agencies and organizations
 - Is multidisciplinary with diverse and disparate data
 - Has research and (to) operations
 - Is a challenging science
- **Curation and availability of operational models**
 - Difficult to obtain operational model outputs from SWPC
 - not their mission to curate data
 - Cannot obtain operational model outputs from NCEI
 - their mission is to curate observations
- **Curation and availability of research models**
 - CCMC is required to store models and resulting data
 - May not be the most efficient use of resources (e.g., if \$compute < \$storage)
 - Potentially better to store model's initial and boundary conditions with subset of model outputs.
 - Can re-run model with stored IC/BC etc.
 - Difficult to obtain from academic and laboratory (local infrastructure – lack of access)
- **Versioning**
 - Multiple versions of models exist at multiple locations
 - Where is the authoritative operational model?
 - Different agencies etc. can have different requirements
 - → Different models/implementations are therefore required
 - Where is the authoritative research model to be continually improved?
 - Is the dev environment set up for CI/CD? No? Where are the software engineers?!
- **Space-based v. ground-based: a specific example**
 - MagPy can be run on HMI magnetograms to predict flares
 - HMI is a space-based scientific instrument
 - GONG are ground-based operational instruments similar to HMI
 - MagPy is much more likely to perform better with HMI
 - Researchers are caught up in “should I focus on research using X data that will likely have better results and garner more scientific interest or focus on research using Y data that is more likely to find support to transition to ops?”
 - Will NASA bake-in operations to Heliophysics missions? \$\$
 - For the research community we need better guidance and documentation on what data we can use!!



Data Opportunity Examples

- Many of the next generation data curation and R20-02R efforts are nascent and going on in parallel ... Perfect!!!:
 - NASA HPDE/HPCloud
 - SWPC Testbed
 - Space Force SET4D
 - Air Force UDL
 - NOAA AI
- Specific: Space Wx event associations
 - Create scientific crowd sourced event catalogs w/ versioning!
 - Use APIs and tools (e.g., HAPI) to identify other assoc. events
 - Streamline data normalization and data fusion → **data discovery**
- Technology: Dask, Xarray, Bokeh, kerchunk, cloud
- We are a small community
 - We can and are organizing!
 - If we continue to be strategic, tactical, vigilant, creative, and secure more resources, we will develop the solutions needed



Flare, SEP, shock, CME event associations

ARs → flares → CMEs → radio bursts → SEPs → IMF → shocks → solar wind structures → magnetosphere → atmosphere → ionosphere → geoelectric
→ **technology impacts (e.g., satellite charging, human health, GICs, communications)**



Acronyms

- CCMC – Community Coordinated Modeling Center
- CI/CD – Continuous Integration and Continuous Deployment
- DevSecOps – Development, Security and Operations
- HAPI – Heliophysics Application Programming Interface
- NCEI – National Centers for Environmental Information
- NESDIS – National Environmental Science, Data, and Information Services
- O2R – operations to research
- R2O – research to operations
- SWPC – Space Weather Prediction Center



1957 Topps Space Cards #78