Article:
Seaton, D. orcid.org/0000-0002-0494-2025, Caspi, A. orcid.org/0000-0001-8702-8273, Casini, R. orcid.org/0000-0001-6990-513X et al. (61 more authors) (2023) Improving multi-dimensional data formats, access, and assimilation tools for the twenty-first century. Bulletin of the AAS, 55 (3). ISSN 2330-9458

https://doi.org/10.3847/25c2cfeb.6d8ecdc1

A White Paper Submitted to the Solar & Space Physics (Heliophysics) Decadal Survey — 2024–2033

Improving Multi-Dimensional Data Formats, Access, and Assimilation Tools for the Twenty-First Century

Primary Author: Daniel B. Seaton[1], **Core Team Co-Authors:** Amir Caspi[1], Robert Casini[2], Cooper Downs[3], Sarah E. Gibson[2], Holly Gilbert[2], Lindsay Glesener[4], Silvina Guidoni[5], J. Marcus Hughes[1], David McKenzie[6], Joseph Plowman[1], Katharine K. Reeves[7], Pascal Saint-Hilaire[8], Albert Y. Shih[9], and Matthew J. West[1]

[1]Southwest Research Institute, Boulder, CO, [2]National Center for Atmospheric Research, Boulder, CO, [3]Predictive Science Inc., San Diego, CA, [4]University of Minnesota, Minneapolis, MN, [5]American University, Washington, DC, [6]NASA Marshall Space Flight Center, Huntsville, AL, [7]Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA, [8]University of California, Berkeley, CA, [9]NASA Goddard Space Flight Center, Greenbelt, MD

Additional Co-Authors: Refer to attached spreadsheet.

Synopsis

Heliophysics image data largely relies on a forty-year-old ecosystem built on the venerable Flexible Image Transport System (FITS) data standard. While many *in situ* measurements use newer standards, they are difficult to integrate with multiple data streams required to develop global understanding. Additionally, most data users still engage with data in much the same way as they did decades ago. However, contemporary missions and models require much more complex support for 3D multi-parameter data, robust data assimilation strategies, and integration of multiple individual data streams required to derive complete physical characterizations of the Sun and Heliospheric plasma environment. In this white paper we highlight some of the 21st century challenges for data frameworks in heliophysics, consider an illustrative case study, and make recommendations for important steps the field can take to modernize its data products and data usage models. Our specific recommendations include:

- Investing in data assimilation capability to drive advanced data-constrained models,
- Investing in new strategies for integrating data across multiple instruments to realize measurements that cannot be produced from single observations,
- Rethinking old data use paradigms to improve user access, develop deep understanding, and decrease barrier to entry for new datasets,
- Investing in research on data formats better suited for multi-dimensional data and cloud-based computing.

1 The Challenge for Heliophysics Data Standards

The Flexible Image Transport System (FITS), the prevailing standard for image-based data products in heliophysics, was developed in the late 1970s, when computerized analysis of electronic image data was just becoming the norm in astronomy. FITS was intended to stop the proliferation of ad hoc formats and provide a simple standard – well suited to storage on magnetic tape – that was both human and machine readable (Wells et al., 1981).

The standard's design was brilliant: engineered simultaneously to be appropriately sized and formatted for tape files, compatible with almost all computers of the era, and accessible even to data users who were unfamiliar with the new format. In so doing, the standard also (preemptively) complied with recommended best practices for archival data formats, which mandate such formats should still be interpretable even in the absence of the computer or software systems for which they were designed (National Research Council, 1995).

Since those early days, FITS has been fully standardized, upgraded, and refined to better carry complex, compressed, multi-dimensional data needed for modern observations and computers, reaching version 4.0 in the present day (IAU FITS Working Group, 2018). FITS has also become the single standard for nearly all heliophysics image data, and the community has developed infrastructure that allows users to search and download data (i.e., the Virtual Solar Observatory Hill et al., 2009) in FITS format from dozens of instruments.

Though FITS is capable of containing multi-dimensional data, its primary purpose at conception was to transmit image data, generally serially, in individual files, for local calibration and analysis. Historically the data use model for heliophysics observations in FITS format has been to obtain individual observations in individual files, transport them to a local file system, and calibrate and analyze them there.

In an era in which data were generally 2D representation of camera output, whether image or spectra, this model was generally workable. However, many contemporary and proposed instruments produce vastly more complex data (e.g., Cheung et al., 2019; Golub et al., 2020), require advanced image processing tools and significant computing resources to calibrate and process (e.g., Winebarger et al., 2019; DeForest, 2017), and produce time-varying, multi-dimensional datasets that are difficult to represent under old paradigms. Increasing data product volume has made local analysis of advanced data less and less feasible, even in spite of improvements in internet bandwidth, necessitating online analysis environments with simplified or degraded representations of quantitative data (Müller et al., 2017). Many data products are not optimized for integration with physical models, reducing their value as model drivers or constraints, and slowing progress on data assimilation in heliophysics compared to other data-rich scientific domains. Integrating data products across multiple instruments is often infeasible, or requires extensive expertise in multiple projects, which few researchers possess.

Since the advent of FITS 40 years ago, heliophysics has evolved from a data-limited research environment with nascent numerical modeling capabilities, to a data-rich one with advanced models with extensive data assimilation needs. There is a critical need for new approaches to data products and data assimilation strategies. Through the development of the *COMPLETE* mission concept (see additional white papers by Caspi et al., 2022a,b), we have identified new strategies for complex, multi-perspective, multi-dimensional data products, and recommendations for investments that could help realize a new vision for heliophysics data for the next century. In this white paper we highlight a case study (Sec. 2) and provide recommendations for these improvements and investments (Sec. 3).

2 Case Study: Data Integration for COMPLETE

The COMPLETE mission concept embodies the need for advanced data products and processing for contemporary missions, where multiple data streams from disparate measurements must be integrated within a unified physical framework. COMPLETE comprises two integrated instrument suites – a comprehensive 3D magnetograph and broadband spectral imager – distributed across multiple spacecraft at differing solar view angles. The comprehensive magnetograph combines surface field measurements from a photospheric magnetograph instrument with magnetic diagnostics in the corona using a Hanle-effect Lyman-α polarization coronagraph (Raouafi et al., 2016). The broadband spectral imager combines observations from γ-rays, X-rays, and EUV along with multi-messenger energetic neutral atom observations to deduce plasma properties in the corona. To integrate these disparate measurements, the instruments are specifically curated and co-optimized to produce mutually compatible observations, which then must be assimilated into an overall 3D data framework within physical context.

Some plasma properties, such as temperature and density, can be deduced from broadband spectral images straightforwardly, using differential emission measure techniques (DEM; e.g., Cheung et al., 2015; Plowman & Caspi, 2020). However, traditional applications of these techniques can only reveal the properties projected into 2D images, and integrated along the line of sight.

Techniques to invert Stokes profiles to determine magnetic fields have existed for decades (Auer et al., 1977) and are now both robust and computationally advanced (Borrero et al., 2011). However, to characterize the corona's magnetic properties, these techniques must be coupled to models that extrapolate or predict the coronal magnetic field. In the absence of additional constraints, solving for the coronal field is a massively underdetermined problem (Plowman, 2021). However, Hanle-effect measurements can provide the missing constraint – if they can be assimilated into the overall data/model framework.

Assimilation of multi-instrument, multi-messenger, multi-dimensional data is part of routine research and forecasting operations in data-rich fields such as atmospheric science and meteorology (see the review by Lahoz & Schneider, 2014, and references therein). While this is new territory for heliophysics, with appropriate investment to adapt existing strategies there is no reason heliophysics cannot achieve the same level of sophistication and success as these other fields. For example, for COMPLETE, we studied a straightforward but powerful method that can integrate magnetic field information and plasma properties to generate a 3D model of key physical parameters in the corona using multi-perspective observations. These reconstructed parameters can then serve as model initial conditions or constraints within more sophisticated simulations.

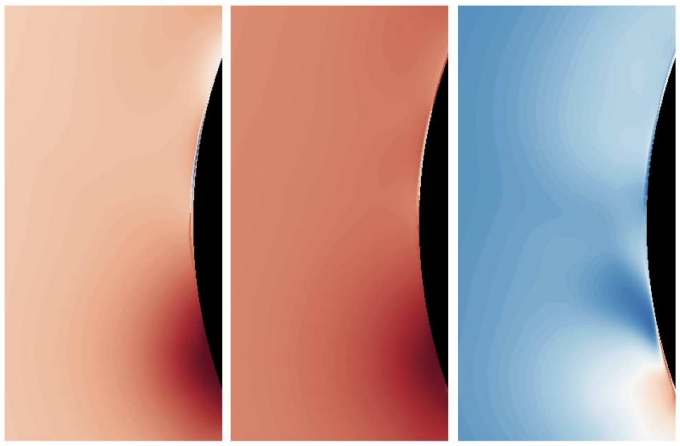

Figure 1: Forward-modeled Hanle-effect polarization through different line-of-sight depths (0.1, 0.2, & 1.0 $R_\odot$, left to right), demonstrating the diagnostic potential of these observations of coronal fields. The underlying model is a Magnetohydrodynamic Algorithm outside a Sphere (MAS) simulation of the corona at the time of the 2017 total solar eclipse (Mikić et al., 2018).

2.1 Reconstructing the 3D Corona

The approach we are developing for COMPLETE uses an extension of the Coronal Reconstruction Onto B-Aligned Regions (CROBAR; Plowman, 2021, 2022) method. CROBAR leverages the fact that the high conductivity of the corona confines plasma to magnetic field lines (the so-called "frozen-flux" condition) to generate 3D reconstructions of plasma distribution within a magnetic extrapolation from photospheric boundary conditions. By coupling the reconstruction technique to data-constrained magnetic extrapolations, DEM tools, and multiple perspectives, it is possible to obtain an accurate 3D reconstruction of the coronal temperature, density, pressure, and magnetic field – and therefore derived and correlated properties like plasma β – within the reconstructed volume.

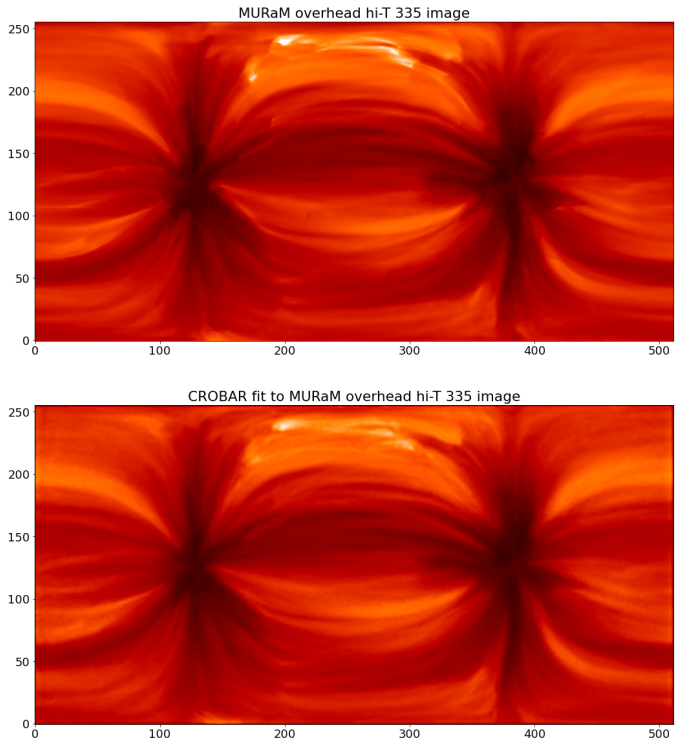

Figure 2: Synthesized image of coronal loops from the MuRAM model (top; Rempel, 2017) compared to an image derived from a 3D CROBAR reconstruction (bottom; Plowman, 2022).

CROBAR presently uses a linear force-free magnetic model tuned by comparison with optically thin emission (e.g., EUV) images. However, its reconstructions can also drive forward models of Hanle effect observables (Gibson et al., 2016, see Fig. 1), which are directly connected to the coronal magnetic field, in much the same way that CROBAR optimizes the magnetic field by comparison with the emission observations. The method already achieves impressive fidelity with EUV observations alone (Fig. 2), and with direct coronal field constraints it will be better still.

Such a reconstruction technique (see Fig. 3), which synthesizes multi-point observations of surface magnetic fields, coronal magnetic field diagnostics, and broadband spectral images, demonstrates, in a simple package, the feasibility of reconstructing key coronal parameters with limited computational overhead. **Reconstructed 3D parameters could then serve as data-driven initial conditions or other constraints for numerical simulations of the corona, and thus provide a straightforward path for a fully integrated data assimilation strategy**.

2.2 A Challenge to Traditional Data Management Strategies

3D reconstructions of many physical parameters, such as this example, present a major opportunity for heliophysics, but also a major challenge. As envisioned by the COMPLETE concept, each reconstruction represents a snapshot in time with a set of high-resolution, multi-dimensional data: three spatial dimensions in spherical coordinates, vector magnetic field, and plasma parameters including temperature, density, and pressure (which themselves may need to be multidimensional, as coronal plasma can be multi-thermal) as well as the original observables. These reconstructions must also include extensive metadata to describe their complex primary data so that researchers

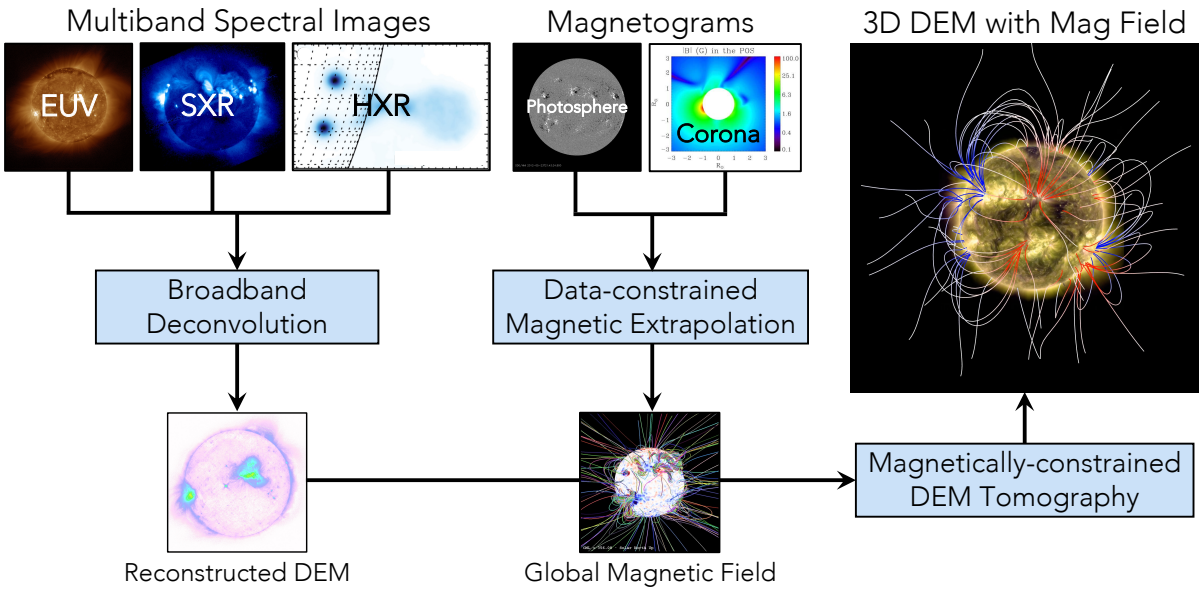

Figure 3: COMPLETE's strategy to integrate multi-faceted data into a unified data-constrained model of conditions in the corona highlights the types of opportunities advanced data assimilation and model strategies may provide.

and their models and analysis tools can interact with it appropriately.

It is unlikely such reconstructions can be efficiently represented within the current limitations of the venerable FITS standard, and new file formats will be needed. Potential candidate standards already exist in other data-rich fields, including:

- Network Common Data Form (netCDF; Unidata, 2021), widely used for atmospheric and in situ space physics data and optimized for multi-dimensional arrays with complex metadata;
- Environmental Systems Research Institute (ESRI) Shapefiles (Environmental Systems Research Institute, Inc., 1998), widely used for vector data in for geographic information system applications;
- Zarr (Zarr, 2022), a young, promising cloud-optimized format, which has seen some use in heliophysics through recent projects at NASA's Frotier Development Lab in 2022.

Each of these has seen use in applications with similar requirements to those of modern heliophysics data, and offers a degree of standardization and community adoption (albeit, within other communities) that could help to facilitate adoption within heliophysics as well. However, more research is required to develop specific requirements necessary to identify a common format that would be appropriate for advanced data such as ours across an entire community.

Likewise, existing analysis frameworks may be insufficient to quantitatively explore these new products. Though tools like jHelioviewer provide good pathfinders for *qualitative* analysis of multi-perspective, multi-band image data, they do not permit deep exploration of quantitative data in 3D. The size of these multidimensional products is likely to upend traditional models of data distribution, potentially necessitating new cloud-based analysis environments, where users can work within the full multi-dimensional data set, extract key measurements, and only download the subset of measurements, plots, or renderings they require locally at the end of their analysis activity. A few such environments already exist, including example implementations using Google's *Colab* notebook-based coding environment (Tamayo & Bellis, 2021) and the Space Radiation Intelligence System (SPRINTS; Engell et al., 2017) project, but much more research and development is required to realize the full-featured analysis environment that will be required for such modern data.

3 Recommendations: Data for 21st Century Science

To continue to support cutting edge science during the coming decade, the heliophysics community must adopt new, innovative thinking about what data products are and how to use them. We have identified four major strategic priorities to address the needs of the data- and model-rich environment developing within the field.

Strategic Priorities

- Invest in data assimilation capability to drive advanced data-constrained models,
- Invest in new strategies for integrating data across multiple instruments to realize measurements that cannot be produced from single observations,
- Rethink old data use paradigms to improve user access, develop deep understanding, and decrease barriers to entry for new datasets,
- Invest in research on data formats better suited for multi-dimensional data and cloud-based computing.

Data Assimilation. Many numerical models in heliophysics use direct measurements as boundary conditions for advanced calculations, but most of these models still lack sufficient constraints to ensure they accurately represent the underlying physical systems. However, with multi-perspective observations, routine measurements of coronal magnetic fields, and advanced data processing techniques, new constraints will become available in the coming decade. Novel approaches are required to ingest these new measurements, along with direct in situ sampling of coronal plasma from missions like Parker Solar Probe, into models that were not developed with such constraints in mind. Interdisciplinary research efforts, particularly in coordination with other fields, such as atmospheric science, with robust support for model data assimilation will be of value. These strategies, and the advanced models they support, will be of value both for basic research and space weather forecasting – strategic planning will allow missions and data products to be developed with these broad applications in mind.

Multi-observation Integration. Traditional models for data products were developed when most data were single, 2D images or spectra, and were intended to be used in isolation from other observations. As more observational capability emerged, researchers developed strategies to deal with complementary observations (e.g., EUV coronal images from AIA and corresponding magnetographs from HMI). A few techniques, such as DEM analysis, exist that can ingest data from multiple observatories, but they are not robust, since the observations themselves were not developed with cross-instrument integration in mind.

New missions in development (such as the COMPLETE mission concept) prioritize this kind of cross-instrument data integration to achieve results that cannot be obtained from single observations alone. Both investments in data integration methods, such as those outlined in Sec. 2 here, and strategic planning, to identify and develop compatible datasets for this purpose, are needed.

An potential pathfinder may be the Polarimeter to Unify the Heliosphere and Corona Small Explorer mission (PUNCH; DeForest et al., 2022), which integrates the observations of four individual imagers into complete, polarimetric observations of the corona and heliosphere between 6–180 $R_\odot$. The instrument and data processing design strategies developed for PUNCH point to both challenges and opportunities to overcome them that will have broad applicability to further

research into multi-instrument measurements. Future missions that make more complex observations – and produce more complex data – will require even more complex strategies. Targeted investments into research on multi-instrument, multi-perspective integrated data are required.

A New Data Use Paradigm. Present-day data analysis models, in which data users acquire and analyze data locally, have their roots in the pre-internet era – far before the development of today's cloud computing strategies. But as data have grown in complexity, this model has led to increased barriers for researchers looking to work with multiple observation sets, both because the data volume can become restrictively large, and because specialized expertise is required to work with new data.

New thinking about how to constitute data products, as well as how and where we work with them, can break down these barriers. Data products that contain derived physical measurements are easier for new users to interpret than direct observations (from which these parameters must be extracted), and do not require the user access complex and computationally expensive processing software to make measurements. Cloud computing solutions are increasingly viable, and permit users to access data without extensive transfers to local systems. Additional work is needed to improve cloud-based visualization capabilities, determine how to manage data egress costs, and develop robust, powerful analysis tools for these environments.

Advanced Data Formats. The new data applications discussed above may result in products or work environments for which current data standards are not optimized. Community-wide investments are required to identify new standards, or improvements to existing standards, to support more advanced, modern data applications. Standards at use in other communities may be applicable or adaptable to the heliophysics community's needs. Extensions of existing standards may also be possible, analogous to the incorporation of Zarr's cloud-ready capabilities into netCDF via the NCZarr format (see "NCZarr Introduction" in Unidata, 2021). Community assessment of the requirements for contemporary data standards and investments to support the development (or re-development) of data formats to meet these requirements are critical.

Formats and analysis environments that leverage the cloud represent a major step forward, but also require community investments to ensure long-term maintenance, stewardship, and appropriate curation. Funding agencies must develop strategies to ensure sustained data access under such a model, as well as resources to support appropriate levels of data egress (e.g., for rendered movies, figures, and subsets of data required for local analysis).

Investments in these strategic priorities, coupled with the development of new observational capabilities for 3D plasma parameters and magnetic fields, can provide keys to true *transformative progress* within heliophysics during the coming decades. By building robust data products that leverage lessons learned from cross-discipline research on data assimilation, heliophysics can finally realize both comprehensive understanding of the Sun-Heliosphere system as a whole, and major gains in our ability to study, forecast, and track drivers of space weather that are critical to long-term risk management in our space-faring, technologically-driven society.

References

Auer, L. H., Heasley, J. N., & House, L. L. 1977, "The determination of vector magnetic fields from Stokes profiles.", *Solar Phys.*, 55, 47, doi: 10.1007/BF00150873

Borrero, J. M., Tomczyk, S., Kubo, M., et al. 2011, "VFISV: Very Fast Inversion of the Stokes Vector for the Helioseismic and Magnetic Imager", *Solar Phys.*, 273, 267, doi: 10.1007/s11207-010-9515-6

Caspi, A., Seaton, D. B., Casini, R., Downs, C., & Gibson, S. E. 2022a, "COMPLETE: A flagship mission for complete understanding of 3D coronal magnetic energy release", A White Paper Submitted to the 2022 Heliophysics Decadal Survey

—. 2022b, "Magnetic Energy Powers the Corona: How We Can Understand its 3D Storage & Release", A White Paper Submitted to the 2022 Heliophysics Decadal Survey

Cheung, M. C. M., Boerner, P., Schrijver, C. J., et al. 2015, "Thermal Diagnostics with the Atmospheric Imaging Assembly on board the Solar Dynamics Observatory: A Validated Method for Differential Emission Measure Inversions", *Astrophys. J.*, 807, 143, doi: 10.1088/0004-637X/807/2/143

Cheung, M. C. M., De Pontieu, B., Martínez-Sykora, J., et al. 2019, "Multi-component Decomposition of Astronomical Spectra by Compressed Sensing", *Astrophys. J.*, 882, 13, doi: 10.3847/1538-4357/ab263d

DeForest, C., Killough, R., Gibson, S., et al. 2022, "Polarimeter to unify the corona and heliosphere (PUNCH): Science, status, and path to flight", in 2022 IEEE Aerospace Conference (AERO), 1–11, doi: 10.1109/AERO53065.2022.9843340

DeForest, C. E. 2017, "Noise-gating to Clean Astrophysical Image Data", *Astrophys. J.*, 838, 155, doi: 10.3847/1538-4357/aa67f1

Engell, A. J., Falconer, D. A., Schuh, M., Loomis, J., & Bissett, D. 2017, "SPRINTS: A Framework for Solar-Driven Event Forecasting and Research", *Space Weather*, 15, 1321, doi: 10.1002/2017SW001660

Environmental Systems Research Institute, Inc. 1998, "Esri shapefile technical description", Tech. rep. https://support.esri.com/en/white-paper/279

Gibson, S., Kucera, T., White, S., et al. 2016, "FORWARD: A toolset for multiwavelength coronal magnetometry", *Frontiers in Astronomy and Space Sciences*, 3, 8, doi: 10.3389/fspas.2016.00008

Golub, L., Cheimets, P., DeLuca, E. E., et al. 2020, "EUV imaging and spectroscopy for improved space weather forecasting", *Journal of Space Weather and Space Climate*, 10, 37, doi: 10.1051/swsc/2020040

Hill, F., Martens, P., Yoshimura, K., et al. 2009, "The Virtual Solar Observatory—A Resource for International Heliophysics Research", *Earth Moon and Planets*, 104, 315, doi: 10.1007/s11038-008-9274-7

IAU FITS Working Group. 2018, "Definition of the flexible image transport system v4.0", Tech. rep. http://fits.gsfc.nasa.gov/iaufwg/

Lahoz, W. A., & Schneider, P. 2014, "Data assimilation: making sense of earth observation", *Frontiers in Environmental Science*, 2, doi: 10.3389/fenvs.2014.00016

Mikić, , Z., Downs, C., et al. 2018, "Predicting the corona for the 21 August 2017 total solar eclipse", , 2, 913, doi: 10.1038/s41550-018-0562-5

Müller, D., Nicula, B., Felix, S., et al. 2017, "JHelioviewer. Time-dependent 3D visualisation of solar and heliospheric data", *Astron. Astrophys.*, 606, A10, doi: 10.1051/0004-6361/201730893

National Research Council. 1995, Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources (Washington, DC: The National Academies Press), doi: 10.17226/4871

Plowman, J. 2021, "Three-dimensional Reconstruction of Coronal Plasma Properties from a Single Perspective", *Astrophys. J.*, 922, 109, doi: 10.3847/1538-4357/ac2664

—. 2022, "Validation & Testing of the CROBAR 3D Coronal Reconstruction Method with a MURaM simulation", arXiv e-prints, arXiv:2209.01753. https://arxiv.org/abs/2209.01753

Plowman, J., & Caspi, A. 2020, "A Fast, Simple, Robust Algorithm for Coronal Temperature Reconstruction", *Astrophys. J.*, 905, 17, doi: 10.3847/1538-4357/abc260

Raouafi, N. E., Riley, P., Gibson, S., Fineschi, S., & Solanki, S. K. 2016, "Diagnostics of Coronal Magnetic Fields Through the Hanle Effect in UV and IR Lines", *Frontiers in Astronomy and Space Sciences*, 3, 20, doi: 10.3389/fspas.2016.00020

Rempel, M. 2017, "Extension of the MURaM Radiative MHD Code for Coronal Simulations", *Astrophys. J.*, 834, 10, doi: 10.3847/1538-4357/834/1/10

Tamayo, G., & Bellis, M. 2021, "IceCube and SDSS data visualization in the cloud with Google's Colab environment", in American Astronomical Society Meeting Abstracts, Vol. 53, American Astronomical Society Meeting Abstracts, 525.03

Unidata. 2021, "netcdf users guide, v1.1", Tech. rep., doi: 10.5065/D6H70CW6

Wells, D. C., Greisen, E. W., & Harten, R. H. 1981, "FITS - a Flexible Image Transport System", *Astron. Astrophys. Suppl.*, 44, 363

Winebarger, A. R., Weber, M., Bethge, C., et al. 2019, "Unfolding Overlapped Slitless Imaging Spectrometer Data for Extended Sources", *Astrophys. J.*, 882, 12, doi: 10.3847/1538-4357/ab21db

Zarr. 2022, Zarr Project. https://zarr.readthedocs.io/