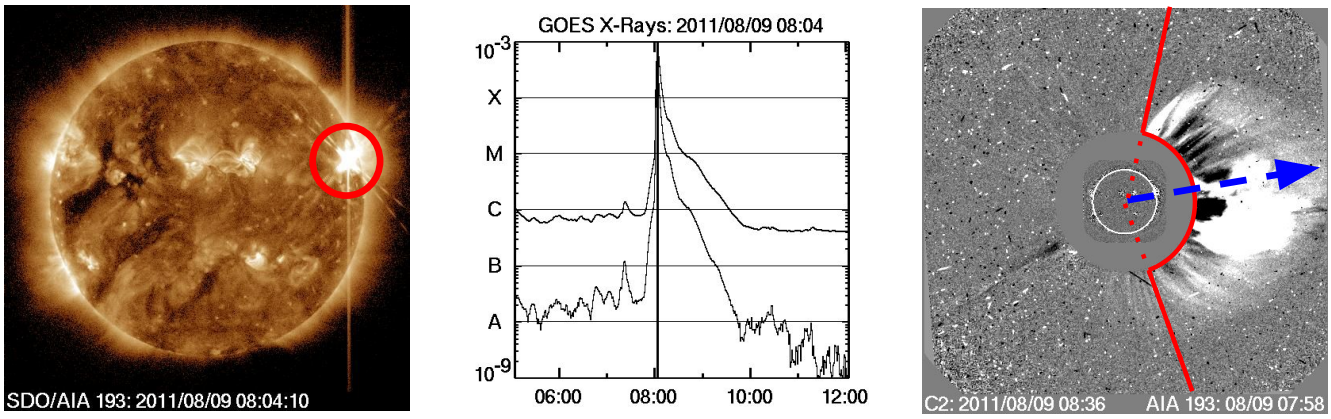

(a) (b) (c)

Figure 1: (a) An SDO/AIA image (193Å) capturing the X6.9 solar flare occurred at 2011-08-09 08:04 (red circle). (b) GOES X-ray flux time series corresponding to the X6.9 flare. (c) The associated CME is observed with a SOHO/LASCO coronograph. Additional red arcs and lines show the width of the CME (~ $130deg$, while the blue line shows the principle angle (~ $280deg$ counterclockwise from the Sun's North pole).

1 GOES FLARES

The flare dataset herein includes both Michelson-Doppler Imager (MDI) magnetograms from SOHO [10] and Helioseismic and Magnetic Imager (HMI) magnetograms from SDO [11]. GOES flares are obtained using the Sunpy GOES module [13], originating from the National Oceanic and Atmospheric Administration (NOAA) National Geophysical Data Center (NGDC). The NOAA/GOES observations measure solar disk-integrated fluxes between $0.1 - 0.8$ nm using the X-ray Sensors (XRS). When a sudden, yet persistent, X-ray flux increase is detected, the event is flagged as a likely flare. A manual review is performed by NOAA forecasters to produce the final NOAA flare list. The GOES flares include three temporal attributes – start, peak, and end time of the flare - and two spatial attributes, namely an explicit coordinate (latitude and longitude) and NOAA active region (AR) number, which can be used to derive the position of each event. NOAA uses data from the Solar X-ray Imager (SXI), replaced recently with Solar Ultraviolet Imager (SUVI), on the same GOES satellites to capture full-disk images with a cadence of one minute in filter bands ranging from 0.6 to 6 nm, as well as other data sources, to pin-point each flare location. In Fig. 1.(a), we demonstrate the SDO/AIA image (193 Åchannel) capturing the X6.9 flare that occurred at 2011-08-09 08:05 UT, which occurred close to the West limb of the Sun on the northern hemisphere (coordinates N17W69). Additionally in Fig. 1.(b), we show the associated CME observed within an SOHO/LASCO coronograph.

The spatiotemporal information on flares allows NOAA's Space Weather Prediction Center (SWPC) to co-locate the active region responsible for the flare. Nonetheless, the GOES catalog is not perfect: the locations and NOAA active region numbers are missing for many B-, C- and even a few M-class flares. We, therefore, integrated the centroid locations of NOAA ARs to the GOES flare reports without an explicit spatial location (i.e., only the NOAA AR numbers are listed), and later cross-checked these locations with two independent feature reporting modules: SSW Latest Events and Hinode-XRT Flare Catalog [9].

1.1 SSW Latest Events Flares

SSW Latest Events is an Interactive Data Language (IDL) SolarSoft module developed by the Lockheed Martin Solar and Astrophysical Laboratory (LMSAL), that uses SDO/AIA images to detect the locations of the flares. Flare location assessment is performed by two AIA images (one before the flare peak time and another during or shortly after the flare peak time). The SSW Latest Events module generally provides better spatial accuracy; however, we have performed manual web scraping of SSW Latest Events from their online servers due to data being unavailable through AIA downtimes.

1.1.1 EIT Flares. Similar to SSW's Latest Events detected from AIA images, we also pinpointed the locations of a subset of SOHO-era flares using the 195 Angstrom EIT images. Due to the relatively low temporal resolution of EIT images (which was revised from two images per day to 12 minutes from mid-1998 until 2010), we observe significant data gaps in capturing much of the flaring phenomenon. In such cases, we detected and verified the locations of SOHO-era flares of GOES class >C1.0 for those existing available data.

2 CME SOURCES

The CME information has been obtained by the SOHO Large Angle Spectroscopic Coronagraph (LASCO) CME Catalog [6]. The catalog provides important CME characteristics such as CME's first appearance time in C2, position angle (central and measurement), and linear speed. Using this data, we have paired associated SEPs and the flares with NOAA AR information to connect events to their solar source. We include an additional metadata feature from the delivered dataset of the CMEs' launch time, besides the CMEs' first appearance time reported elsewhere. The CME launch time is obtained by extrapolating the height–time CME diagram used to infer the linear CME speed to zero height, indicating the time the CME progenitor started ascending undetectable, against the solar disk, in the Sun's low atmosphere.

The inferred time difference between the CME launch and first appearance times ranges roughly between a few tens of minutes and a couple of hours. This difference depends mainly on the speed of the CME and the height of the first detection. By virtue of the LASCO C2 coronagraph (which shows roughly the area between 1.5 to 6 solar radii), the first detection may occur within a few ($\leq$5) solar radii or higher. CME velocities range between a few hundred and >3000 km/s. We demonstrate a CME in Fig. 1.(c), which is associated to the X6.9 flare of Fig.1.(a). Two key spatial characteristics of CMEs are the principal angle and the angular width. The width of the CME is the angular spatial span between the lines connecting the two ends of the CME footprint relative to the center of the Sun (shown as the red lines in Fig. 1. The principal angle is the projected central angle of the CME measured in degrees from the North pole of the Sun. Note that while these spatial characteristics are often reported as one value per CME in each CME catalogs, they are time-evolving characteristics that fluctuate over the lifespan of CMEs.

The CME launch time is needed because it helps put to rest at least some paradoxical situations where the SEP is detected at L1 before the (thought as) source CME is registered in the near-Sun sky. For regions of good connectivity (i.e., the western solar hemisphere), accelerated SEPs can be at L1 within a few tens of minutes, with a minimum of ~20 minutes. If the CME needs several tens of minutes to appear and be detected, then the SEP event may be detected at L1 before the source CME is detected in the Sun. The true CME launch time must be before the SEP detection. This said the fit used to obtain the height-time diagram is only linear, while we know that CMEs, depending on their initial conditions, tend to accelerate and then decelerate in their initial stages of propagation (see, e.g., [14] and references therein). In several cases, therefore, the fit is poor and generally considered unreliable.

Having to focus on solar active regions, consequently, we also need diagnostic metadata that predominantly relies on photospheric magnetograms. The dataset delivered herein includes both Michelson-Doppler Imager (MDI) magnetograms from SOHO [10] and Helioseismic and Magnetic Imager (HMI) magnetograms from SDO [11]. Metadata from SDO/HMI magnetograms has already been achieved via the Space Weather Analytics for Solar Flares (SWAN-SF) flare benchmark dataset [1] that includes multivariate timeseries comprising 24 magnetic field parameters for each Space Weather HMI Active Region Patch (SHARP) [2] cutout of the SDO era. For the SOHO era, we use results of the automated Active Region Identification Algorithm (ARIA) [8] [5] that provides a data product similar to SHARP cutouts.

2.1 Active Region Data Sources

2.1.1 NOAA Active Region. NOAA Active Region data plays a crucial role in our data integration methodology. We use NOAA active region (AR) locations to augment unknown or not-reported flare locations. More importantly, the integration of magnetic field parameters, flares, and events is centered around the associated NOAA AR numbers. Each HARPNUM (identifier of HMI Active Region Patch [HARP] and SHARP data series) is associated with zero, one or more NOAA AR numbers. This said, we have identified two issues with NOAA ARs: (1) we have found outliers in NOAA AR daily reports, which showed unexpected location changes and (2) we have identified a few faulty associations (i.e., some NOAA AR numbers not associated with a HARPNUM when they were supposed to or vice versa). Therefore, we provide a manually curated NOAA Active Region list covering 1997 to 2018, together with an updated NOAA AR number to HARPNUM associations.

2.1.2 HMI Active Region Patches. HARPs are identified from HMI LOS magnetograms and later reorganized as definitive series. The HMI pipeline analysis code automatically detects each active pixel in LOS magnetograms and identifies each patch on the full-disk imagery that encompasses an active region. These patches are tracked to create an evolving region trajectory of one or more active regions as they cross the disk. The HARP data series provides the geometrical information to follow the evolving region across the disk. The SHARP data series additionally provides synoptic, HARP-integrated space weather indices for each region at each time step. We use the 12-minute cadence, definitive SHARP series, where the vector magnetic field is remapped to a Lambert Cylindrical Equal-Area projection [12] and decomposed into B_radial, B_phi, and B_theta components. As discussed in the Introduction, however, to achieve a homogeneous, SEP-centered treatment of both SOHO- and SDO-era datasets, we are prepared to use only metadata inferred by the LOS or the radial component of the photospheric magnetic field. Using our designed magnetic field extraction module, we created metadata (inferred by the LOS or the radial component of the photospheric magnetic field) and stored it in a multivariate time series dataset as well as the active-region cutouts of the SOHO era.

2.1.3 ARIA Patches. ARIAPs [3] was developed to cover needs for previous statistical analysis of solar transient activity, where we needed a statistically significant sample of SOHO-era active regions [8]. Given its demonstrated practicality and computing efficiency, it has been used multiple times since then (see, for example, [4]) and for different full-disk photospheric magnetograms, such as those of the Synoptic Optical Long-Term Investigations of the Sun (SOLIS) Vector Spectromagnetograpn (VSM; [7] - ARIA application in [3]). A version of ARIA is also constantly at work for SDO/HMI full-disk LOS magnetograms and is a central tool for ESA's Athens Effective Solar Flare Forecasting (A-EFFort) service, operating since late 2015.

ARIA processes LOS full-disk magnetograms by pattern recognition of accumulations of enhanced magnetic flux in the solar photosphere, that are likely representing active regions. Following the identification of patches of enhanced flux, ARIA generates intensity-weighted centroids for each and moves away from these centroids in an annular fashion, in steps of typical supergranular diameters (i.e., ~30" in the Sun's surface). The outer annulus will include pretty much the entire magnetic flux of the patch and we assign one more supergranular diameter for safety. The prescribed square of this outer circle is the extracted active region cutout. However, for a patch to qualify as an active region, externally imposed tolerance limits of flux imbalance, area, location, and east-west orientation are used. Should a patch qualify, we inspect NOAA SWPC's Solar Region Summary (SRS) and identify NOAA AR number(s) included in the patch. The number closer to the intensity-weighted centroid will be the assigned NOAA AR number of the ARIA patch. If no NOAA AR number exists within an eligible ARIA patch, the patch is named after the heliographic coordinates of its centroid.

The ARIA output is the cutout magnetogram in three versions (LOS, cosine-corrected radial on the projected image plane, and the cosine-corrected radial on the de-projected heliographic plane) along with heliographic latitude and central meridian distance (CMD) for each point in the field of view and global pointing information. ARIA works in IDL and provides both visual results (images) and the respective data in IDL SAVE and FITS formats. Should a patch qualify, we inspect NOAA SWPC's Solar Region Summary (SRS) and identify NOAA AR number(s) included in the patch. The number closer to the intensity-weighted centroid will be the assigned NOAA AR number of the ARIA patch. If no NOAA AR number exists within an eligible ARIA patch, the patch is named after the heliographic coordinates of its centroid.

REFERENCES

[1] Rafal A. Angryk, Petrus C. Martens, Berkay Aydin, Dustin Kempton, Sushant S. Mahajan, Sunitha Basodi, Azim Ahmadzadeh, Xumin Cai, Soukaina Filali Boubrahimi, Shah Muhammad Hamdi, Michael A. Schuh, and Manolis K. Georgoulis. 2020. Multivariate time series dataset for space weather data analytics. *Scientific Data* 7, 1 (July 2020). https://doi.org/10.1038/s41597-020-0548-x

[2] M. G. Bobra, X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, and K. D. Leka. 2014. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-Weather HMI Active Region Patches. *Solar Physics* 289, 9 (April 2014), 3549–3578. https://doi.org/10.1007/s11207-014-0529-3

[3] Manolis K. Georgoulis. 2008. Magnetic complexity in eruptive solar active regions and associated eruption parameters. *Geophysical Research Letters* 35, 6 (Jan. 2008). https://doi.org/10.1029/2007gl032040

[4] Manolis K. Georgoulis. 2011. Are Solar Active Regions with Major Flares More Fractal, Multifractal, or Turbulent Than Others? *Solar Physics* 276, 1-2 (Jan. 2011), 161–181. https://doi.org/10.1007/s11207-010-9705-2

[5] M. K. Georgoulis, N. E. Raouafi, and C. J Henney. 2007. Automatic Active-Region Identification and Azimuth Disambiguation of the SOLIS/VSM Full-Disk Vector Magnetograms. arXiv:0706.4444 [astro-ph]

[6] N. Gopalswamy, S. Yashiro, G. Michalek, G. Stenborg, A. Vourlidas, S. Freeland, and R. Howard. 2009. The SOHO/LASCO CME Catalog. *Earth, Moon, and Planets* 104, 1-4 (Jan. 2009), 295–313. https://doi.org/10.1007/s11038-008-9282-7

[7] C. J. Henney, C. U. Keller, and J. W. Harvey. 2006. SOLIS-VSM Solar Vector Magnetograms. arXiv:astro-ph/0612584 [astro-ph]

[8] B. J. LaBonte, M. K. Georgoulis, and D. M. Rust. 2007. Survey of Magnetic Helicity Injection in Regions Producing X-Class Flares. *The Astrophysical Journal* 671, 1 (Dec. 2007), 955–963. https://doi.org/10.1086/522682

[9] Ryan O Milligan and Jack Ireland. 2018. On the performance of multi-instrument solar flare observations during solar cycle 24. *Solar Physics* 293, 2 (2018), 18.

[10] P. H. Scherrer, R. S. Bogart, R. I. Bush, J. T. Hoeksema, A. G. Kosovichev, J. Schou, W. Rosenberg, L. Springer, T. D. Tarbell, A. Title, C. J. Wolfson, and I. Zayer and. 1995. The Solar Oscillations Investigation - Michelson Doppler Imager. *Solar Physics* 162, 1-2 (Dec. 1995), 129–188. https://doi.org/10.1007/bf00733429

[11] P. H. Scherrer, J. Schou, R. I. Bush, A. G. Kosovichev, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, J. Zhao, A. M. Title, C. J. Schrijver, T. D. Tarbell, and S. Tomczyk. 2011. The Helioseismic and Magnetic Imager (HMI) Investigation for the Solar Dynamics Observatory (SDO). *Solar Physics* 275, 1-2 (Oct. 2011), 207–227. https://doi.org/10.1007/s11207-011-9834-2

[12] John P. Snyder. 1987. Map projections: A working manual. https://doi.org/10.3133/pp1395

[13] The SunPy Community, Will T. Barnes, Monica G. Bobra, Steven D. Christe, Nabil Freij, Laura A. Hayes, Jack Ireland, Stuart Mumford, David Perez-Suarez, Daniel F. Ryan, Albert Y. Shih, Prateek Chanda, Kolja Glogowski, Russell Hewett, V. Keith Hughitt, Andrew Hill, Kaustubh Hiware, Andrew Inglis, Michael S. F. Kirk, Sudarshan Konge, James Paul Mason, Shane Anthony Maloney, Sophie A. Murray, Asish Panda, Jongyeob Park, Tiago M. D. Pereira, Kevin Reardon, Sabrina Savage, Brigitta M. Sipőcz, David Stansby, Yash Jain, Garrison Taylor, Tannmay Yadav, Rajul, and Trung Kien Dang. 2020. The SunPy Project: Open Source Development and Status of the Version 1.0 Core Package. *The Astrophysical Journal* 890 (2020), 68–. Issue 1. https://doi.org/10.3847/1538-4357/ab4f7a

[14] David F. Webb and Timothy A. Howard. 2012. Coronal Mass Ejections: Observations. *Living Reviews in Solar Physics* 9 (2012). https://doi.org/10.12942/lrsp-2012-3