Space Weather

RESEARCH ARTICLE
10.1029/2023SW003765

Key Points:

- Machine learning oriented data set for predicting the occurance and properties of solar energetic particle events
- Multivariate remote sensing and in-situ observations
- Continuous data set spanning several decades

Correspondence to:

K. Moreland,
kimberly.moreland@noaa.gov

Citation:

Moreland, K., Dayeh, M. A., Bain, H. M., Chatterjee, S., Muñoz-Jaramillo, A., & Hart, S. T. (2024). MEMPSEP-III. A machine learning-oriented multivariate data set for forecasting the occurrence and properties of solar energetic particle events using a multivariate ensemble approach. *Space Weather*, *22*, e2023SW003765. https://doi.org/10.1029/2023SW003765

Received 19 OCT 2023
Accepted 11 APR 2024

Author Contributions:

Conceptualization: Kimberly Moreland, Maher A. Dayeh, Hazel M. Bain, Subhamoy Chatterjee, Andrés Muñoz-Jaramillo
Data curation: Kimberly Moreland, Maher A. Dayeh, Hazel M. Bain, Subhamoy Chatterjee, Samuel T. Hart
Formal analysis: Kimberly Moreland, Maher A. Dayeh, Subhamoy Chatterjee
Funding acquisition: Maher A. Dayeh
Investigation: Kimberly Moreland, Maher A. Dayeh, Hazel M. Bain, Subhamoy Chatterjee, Andrés Muñoz-Jaramillo
Methodology: Kimberly Moreland, Maher A. Dayeh, Hazel M. Bain, Subhamoy Chatterjee, Samuel T. Hart
Project administration: Kimberly Moreland
Resources: Maher A. Dayeh

MEMPSEP-III. A Machine Learning-Oriented Multivariate Data Set for Forecasting the Occurrence and Properties of Solar Energetic Particle Events Using a Multivariate Ensemble Approach

Kimberly Moreland[1,2], Maher A. Dayeh[1,2], Hazel M. Bain[3,4], Subhamoy Chatterjee[5], Andrés Muñoz-Jaramillo[5], and Samuel T. Hart[1,2]

[1]The University of Texas at San Antonio, San Antonio, TX, USA, [2]Southwest Research Institute, San Antonio, TX, USA, [3]Cooperative Institute for Research in Environmental Sciences, University of Boulder, Boulder, CO, USA, [4]Space Weather Prediction Center, NOAA, Boulder, CO, USA, [5]Southwest Research Institute, Boulder, CO, USA

Abstract We introduce a new multivariate data set that utilizes multiple spacecraft collecting in-situ and remote sensing heliospheric measurements shown to be linked to physical processes responsible for generating solar energetic particles (SEPs). Using the Geostationary Operational Environmental Satellites (GOES) flare event list from Solar Cycle (SC) 23 and part of SC 24 (1998–2013), we identify 252 solar events (>C-class flares) that produce SEPs and 17,542 events that do not. For each identified event, we acquire the local plasma properties at 1 au, such as energetic proton and electron data, upstream solar wind conditions, and the interplanetary magnetic field vector quantities using various instruments onboard GOES and the Advanced Composition Explorer spacecraft. We also collect remote sensing data from instruments onboard the Solar Dynamic Observatory, Solar and Heliospheric Observatory, and the *Wind* solar radio instrument WAVES. The data set is designed to allow for variations of the inputs and feature sets for machine learning (ML) in heliophysics and has a specific purpose for forecasting the occurrence of SEP events and their subsequent properties. This paper describes a data set created from multiple publicly available observation sources that is validated, cleaned, and carefully curated for our ML pipeline. The data set has been used to drive the newly-developed Multivariate Ensemble of Models for Probabilistic Forecast of SEPs (MEMPSEP; see MEMPSEP-I (Chatterjee et al., 2024, https://doi.org/10.1029/2023SW003568) and MEMPSEP-II (Dayeh et al., 2024, https://doi.org/10.1029/2023SW003697) for accompanying papers).

Plain Language Summary We present a new data set that uses observations from multiple spacecraft observing the Sun and the interplanetary space around it. This data is connected to the processes that create solar energetic particles (SEPs). SEP events pose threats to both astronauts and assets in space. The data set contains 252 solar flare events that caused SEPs and 17,542 that do not. For each event, we gather information about the local space environment around the sun, such as energetic protons and electrons, the conditions of the solar wind, the magnetic field, and remote solar imaging data. We use instruments from NOAA's Geostationary Operational Environmental Satellites (GOES) and the Advanced Composition Explorer spacecraft, as well as data from the Solar Dynamic Observatory, the Solar and Heliospheric Observatory, and the Wind solar radio instrument WAVES. This data set is designed to be used in machine learning (ML), with a focus on predicting the occurrence and properties of SEP events. We detail each observation obtained from publicly available sources, and the data treatment processes used to validate the reliability and usefulness for ML applications.

1. Introduction

Solar energetic particles are high-energy particles associated with two main types of solar activity: solar flares (SFs) and Coronal Mass Ejections (CMEs) (Desai & Giacalone, 2016; Reames, 2013). Solar energetic particle (SEP) flux enhancements last from tens of minutes to days and include protons, electrons, and heavier ions. The proton particle energies can sometimes reach giga electron volts (Reames, 2001). Particles of this energy range can negatively affect technological assets in space, (Horne et al., 2013; Maurer et al., 2017) cause high dose radiation exposure of astronauts and even affect passengers and crews on polar route commercial airline flights (Chancellor et al., 2014; Onorato et al., 2020).

Solar flares originate in the lower solar corona and chromosphere in regions of complex magnetic fields, also known as active regions (ARs). Flares are easily recognized by their large enhancements in the Extreme Ultraviolet (EUV) and X-ray frequency bands. X-rays occur due to collisions between decelerating closed-loop particles and the underlying plasma (Galloway et al., 2010). The Solar Dynamics Observatory (SDO) observes SFs at 11 passbands of differing wavelengths. The 171 Å band allows for investigating the flare intensity, location, and evolution. X-ray flares are often accompanied by Type III radio bursts (i.e., Figure 3). Type III radio bursts are radio waves associated with electrons accelerated at solar magnetic reconnection sites at or near the flaring region that travel along open magnetic field lines through the upper corona and into interplanetary (IP) space (Cairns et al., 2018). These electrons are remotely observed in radio spectrograms as quick radio bursts sweeping from high to low frequencies. While single type III bursts are typically associated with impulsive flares (that are significantly less energetic), the relationship between type III radio bursts and large SEP events is more complex. Some SEP events are associated with multiple decametric/hectometric type III radio bursts that occur between flare onset and flare peak, while other SEP events have a single, large metric type III radio burst (see Klein 2021 and references therein). In addition to flares, ARs can also produce CMEs.

Some ARs develop a build-up of mass along a closed magnetic field line that eventually erupts, releasing a CME. CMEs appear as expanding loops or bubbles and are often seen in visible light coronagraphs (e.g., Large Angle and Spectromic Coronagraph (LASCO)). LASCO records white light images of the solar corona from 2.0 up to 30 solar radii and spectral images of the solar corona from 2.0 to 3.0 solar radii. Coronagraph images make it possible to view the field topology of the corona and its changes, as well as the evolution of CMEs as they travel outward from the sun. CMEs, moving much faster than the preceding material, act as the driver of IP shocks. One indicator of a shock occurrence is a Type II radio burst (E. Cliver et al., 1986; Mäkelä et al., 2011), which is generally observed as a slower high to low frequency sweep and is associated with electrons accelerated by the outward propagating shocks. At the time of the shock passage from the observer, the local particle flux may be enhanced from suprathermal energies to tens of MeV/nucleon. This sudden, short-lived enhancement is known as an energetic storm particle (ESP) event and is only seen at lower energies (Dayeh et al., 2018; Desai et al., 2016; Moreland et al., 2023). The diffusive shock acceleration (DSA) mechanism is the primary candidate for ESP acceleration at shocks. It involves both shock-drift acceleration at quasi-perpendicular shocks (Decker, 1981) and the first-order Fermi mechanism at quasi-parallel shocks (Lee, 1983). Ideally, DSA theory provides an explanation for several aspects of ESP observations and predicts certain particle profile behaviors (e.g., spectral index). However, numerous physical processes such as particle transport, localized turbulence, magnetic connectivity along the shock, shock geometry, and other factors can affect particle transport and acceleration (Giacalone & Neugebauer, 2008; Mostafavi et al., 2018; Zank et al., 2010).

Understanding the physical processes behind SEPs and predicting SEP events as well as their properties (i.e., peak flux, energies, onset, duration, etc.) using true probabilities is becoming exceedingly important. Fortunately, the heliophysics and space weather communities now benefit from large amounts of free and publicly available remote sensing and in situ observations collected over decades. To a greater extent, we can take advantage of the significant advances in computing power, open-source software, and validated algorithms that make for a perfect combination for ML applications (Camporeale, 2019). There are a variety of available data sets in the community such as Space-Weather Helioseismic and Magnetic Imager (HMI) Active Region Patches (SHARPs) and Space-Weather Michelson Doppler Imager (MDI) Active Region Patches (SMARPs) (Bobra et al., 2021), which are derived from MDI and Solar and Heliospheric Observatory (SoHO) solar surface magnetic field maps to provide a seamless set of maps and keywords describing each AR observed since 1996; a multivariate time series data set available via a web API, extracted from SHARPs data and cross-checked with the NOAA SF catalog that includes 51 SF predictive parameters spanning from 2010 to 2018 (Angryk et al., 2020); and a multitude of SEP event lists (i.e., Crosby et al., 2015; Gopalswamy et al., 2014; Kahler et al., 2017; Papaioannou et al., 2016; Rotti et al., 2022). The data set described in this paper consists of a multitude of both remote and in situ measurements that, as we describe in the sections below, are linked to various SEP properties. Developing such a comprehensive data set enables us to investigate numerous SEP properties and their pre-flare conditions. As a result, we are able to ingest our data set into a newly developed model for forecasting both the occurrence of SEP events as well as their properties. We refer to this model as the "Multivariate Ensemble of Models for Probabilistic Forecast of SEPs" (MEMPSEP; see MEMPSEP-I (Chatterjee et al., 2024) and MEMPSEP-II (Dayeh et al., 2024) for papers).

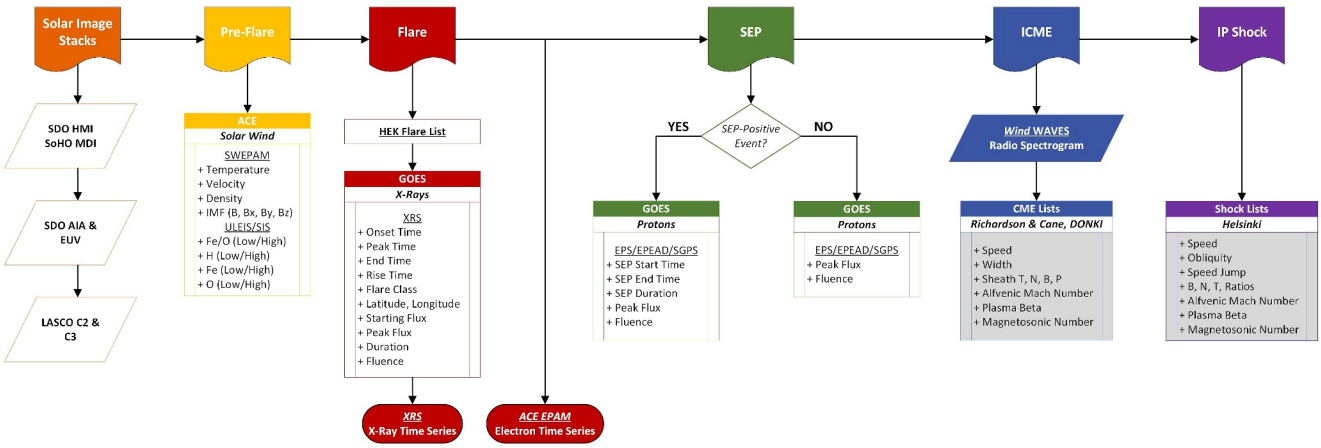

Figure 1. The complete data set flowchart shows all incorporated observations from remote imaging to in-situ measurements. Each section relates to a time frame in the event series: Solar images (pre-flare, flare), solar wind conditions (pre-flare), X-ray properties, and time series (flare). Post-flare properties for the SEP event, Interplanetary Coronal Mass Ejection (ICME) parameters with Wind WAVES radio spectrogram, Interplanetary (IP) shock properties, and published shock lists. We note the observation (italics) and the instrumentation (underlined) used to obtain in-situ data along with each parameter observed or calculated parameter.

When assembling this data set, we consider each physical process independently and select observations that are associated with those processes. To ensure that the data set can be used in forecasting and now-casting, we prioritize data that is made available in near real-time or data that will be available in real-time on future missions. Figure 1 shows the complete set of inputs for the data set, including any calculated parameters and the use of various event lists (see Section 3 for details on event lists). By considering all the solar processes that can potentially alter SEP properties in interplanetary space (i.e., acceleration, transport, diffusion, among others), we create a whole-picture data set that captures each observation's role in determining the occurrence probability and corresponding properties of the resulting SEP events.

2. Instrumentation

Overall, we accumulated data from 13 different instruments over five separate missions. Table 1 summarizes the spacecraft, instrument, measured observations, energy range, and time cadence.

Table 1
Complete Instrumentation Table

Mission	Instrument	Measurement	Energy range	Resolution
SOHO	LASCO	Solar Corona (C2,C3)		1 hr
	MDI	Line of Sight Solar Magnetic Field		6 hr
	EIT	Solar Extreme UV Emissions	171 Å	12 hr
SDO	HMI	Line of Sight Solar Magnetic Field		6 hr
	AIA	Solar Extreme UV Emissions	171 Å	12 hr
Wind	WAVES	Solar Radio Emissions	10 kHz–10 MHz	0.1 s
GOES	EPS	H^+ and e^-	0.6–500 MeV	5 min
	XRS	X-Ray	0.5–8.0 Å	3 s
ACE	EPAM	H^+	0.05–5 MeV	12 s
		e^-	0.0535–0.2325 MeV	12 s
	ULEIS	3 He, 4 He, O, Fe	0.04–10 MeV	1 hr
	SIS	3 He, 4 He, O, Fe	3–170 MeV	1 hr
	SWEPAM	Solar Wind Protons	0.4–4 keV	1 min
	MAG	Local Magnetic Field		16 s

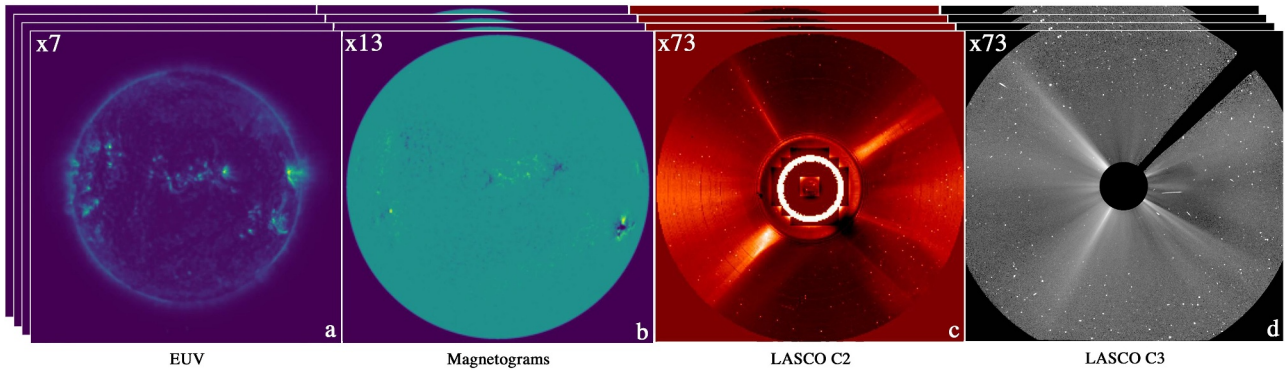

Figure 2. Imaging data is downloaded, and the sequence of images creates a stack that is ingested into the ML model. In this data set, we use a time window beginning at the flare onset to 3 days prior. (a) The EUV stack contains seven images. (b) Magnetogram stack contains 13 images (c), (d) LASCO C2 and C3 stacks contain 73 images each.

2.1. Remote Sensing Observations

To capture the evolution and dynamics of ARs and to help inform ML models, we input remote observations of the solar atmosphere 3 days before the flare onset. We use the Michelson Doppler Imager (MDI; Scherrer et al., 1995) on board SoHO and the HMI (Schou et al., 2012) on board SDO to obtain full line-of-sight magnetograms to investigate the topology of the solar surface and solar atmosphere. Considering the possible memory limitation of graphics processing units (GPUs), we downgrade the resolution of magnetograms to 256 pixels × 256 pixels and collect one magnetogram every 6 hr over 3 days before a flare onset. As a result of variable cadence and the instrument degradation to two images per day in the Extreme Ultraviolet Imaging Telescope (EIT; Delaboudinière et al., 1995) on board the (SoHO; Domingo et al., 1995), we use half-day (12 hr) resolution EUV data from 1996 onward. Starting in 2010, we used EUV data from the Atmospheric Imaging Assembly (AIA; Lemen et al., 2012) on board the SDO (Pesnell et al., 2012) fixed at the same time cadence. To obtain radio observations, we make use of the WAVES instrument Radio Receiver Bands 1 and 2 (RAD1: 20–1,040 kHz, RAD2: 1.075–13.825 MHz) on board *Wind* (Bougeret et al., 1995) stationed at Lagrange Point 1 (L1). Using the 20 channels from RAD1 that contain observed data only (vs. interpolated) and a reduced set of 60 channels from RAD2 with observed data (80 total frequency channels used to reduce data input size) at a down-sampled time of 10 minutes using an anti-aliasing approach (low-pass filtering plus re-sampling), we create a single image of the logarithm of the intensities (see Figure 3). Finally, to observe the changes in the IP space as solar features such as CMEs streaming away from the sun, we obtain hourly images of the solar corona using the LASCO (Brueckner et al., 1995) on board SoHO. Note that we do not fit the CME parameters such as width, propagation speed, duration, etc. In this work, we identify and connect the event parameters to known CMEs via published event lists. An image stack consists of various images corresponding to the instruments' cadence for each remote sensing source. See Section 2 and Figure 2 for more details. HMI, AIA, and MDI data are publicly available via JSOC web portals (http://jsoc.stanford.edu/), and LASCO coronagraph data is obtained using the related SunPy package. Solar imaging offers a large amount of information well-suited to ML algorithms. Keeping in mind data size, we implement data reduction techniques such as under-sampling both in space and time to avoid memory burden and prevent model-crashing.

2.1.1. EUV

We choose the EUV band at 171 Å, which is a wavelength of a strong iron line (Fe IX) at approximately 600,000 K and observes changes in the corona and transition region boundary. This makes it ideal for studying the quiet corona and coronal loops, including fine plasma strands. To create a continuous data set of EUV images (using SoHO and SDO) we download images of overlapping date and time, for example, EIT 171°Å and AIA 171°Å at a cadence of 12 hr over 3 days before flare onset. We filter out the bad images using a "QUALITY" keyword and normalize the images with exposure time. Using the SunPy library, aiapy (Barnes, Bobra, et al., 2020; Barnes, Cheung, et al., 2020), we fetch the correction table to correct for the known SDO/AIA EUV image degradation as described in Galvez et al. (2019). We also perform degridding of the EIT/171 images. We then perform re-projection to homogenize images in terms

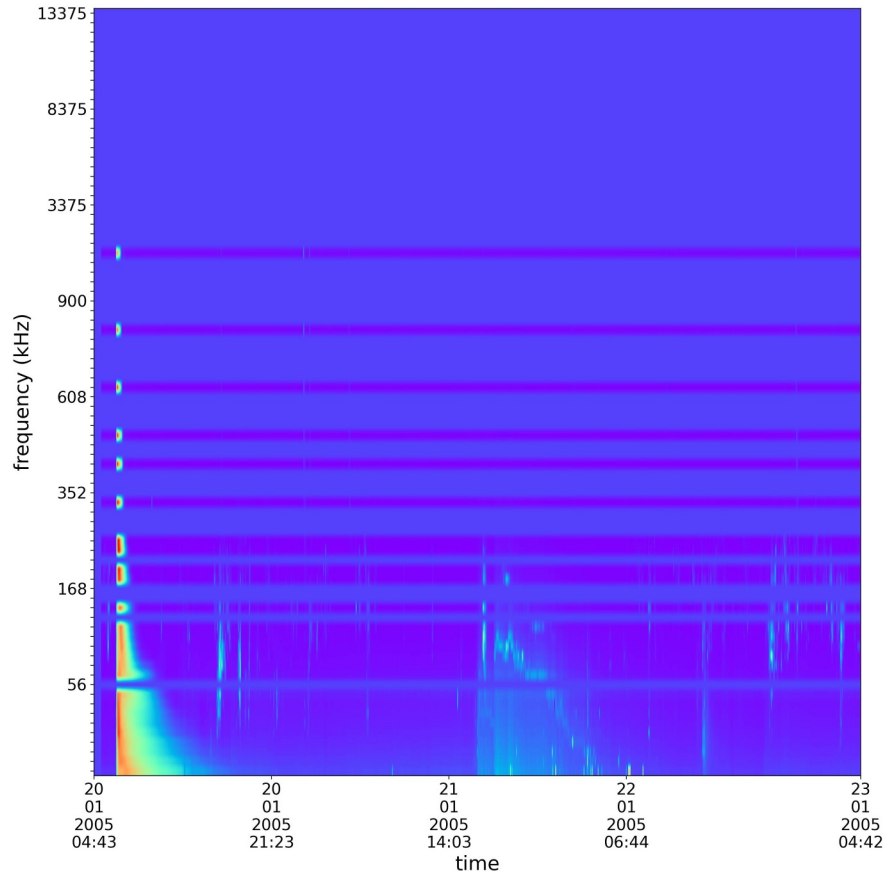

Figure 3. *Wind* WAVES sample event image showing a Type III solar radio burst, evidence of flare accelerated electrons.

of pixel scale, field of view, and point of view. The full-disc images are divided into patches, and the patches are aligned using a template-matching method described in Chatterjee et al. (2023). Finally, we use a low-resolution patch and limb darkening profile (encoding position of the patch pixels w.r.t. disc center) as input and the corresponding high-resolution patch as the output of a Deep Learning (DL) model presented in Chatterjee et al. (2023). We use low-resolution versions (256 pixels × 256 pixels) of the homogenized images (DL model outcome) and, when 50% or more of the images are available in the queried time frame, interpolate for missing frames, creating the EUV data cube of size 256[*pixels*] × 256 [*pixels*] × 7[*frames*] for each flare event. Figure 2a shows a sample EUV image and its corresponding stack size.

2.1.2. Magnetograms

Magnetograms capture the evolution of sunspots and spatial variations in the solar magnetic field properties, especially the surface distribution and polarity of those magnetic fields. SoHO/MDI and SDO/HMI full-disc line-of-sight magnetograms are another imaging input feature of our data set, and a sample event is shown in Figure 2b. Keeping in mind the memory limitation of most GPUs we downgrade the resolution of magnetograms to 256 pixels × 256 pixels. In order to adequately sample the related emergences (i.e., Kutsenko et al., 2019, 2021), we collect one magnetogram every 6 hr over 3 days before the flare onset. This creates a data cube of size 256[*pixels*] × 256[*pixels*] × 13[*frames*]. Note the original size is 1024 × 1024 for MDI and 4096 × 4096 for HMI, with a cadence of 96 min and 45 s, respectively. This image downsampling reduces the computer's memory burden without losing information about the AR evolution. Convolutional Neural Networks (CNN; LeCun et al., 2015), are a specific type of DL algorithm specifically designed for images and uses layers of neurons to identify certain features in images. Despite CNNs being insensitive to timestamps, we maintain regularity by filling 6-hr gaps with nearest-neighbor interpolated magnetograms for each pixel. To homogenize MDI and HMI

magnetograms, we use a conversion factor of 1.3 (MDI = 1.3*HMI) and clip the field strength within (−1,000, 1,000 G). This creates a continuous data set for longer observation periods than a single instrument would provide. Finally, we normalize the field strength (F) using the transformation $\frac{1}{2}(1+\frac{f}{1000})$. So, a pixel value of 0, 0.5, and 1 in the normalized magnetograms represent −1,000, 0, and 1,000 G.

2.1.3. LASCO C2 and C3

Coronographs block the solar disk using an external occulter, revealing coronal features such as coronal streamers and CMEs and their propagation into the interplanetary medium. SOHO has two onboard coronagraphs, C2, which has a 3-degree field of view (1.5–6 solar radii), and C3, which has a 16-degree field of view (3.7–30 solar radii). Sample event images for the LASCO C2 and C3 images can be seen in Figures 2c and 2d, respectively. We download level 0.5 LASCO - C2 and C3 images using JSOC at a cadence of 1 hr over a period of 3 days before flare onset. We apply the SSWIDL (SolarSoft) package to convert level 0.5 to level 1 and apply a normalizing radial graded filter to equalize the contrast of coronal structures with respect to the background at different radial distances. Finally, we down-sample those images to a fixed size of 256 pixels × 256 pixels and perform a nearest neighbor interpolation for each pixel of down-sampled images over the time axis to fill missing frames. This generates a C2 and C3 data cube of size 256[*pixels*] × 256[*pixels*] × 73[*frames*] for each flare event.

2.1.4. Wind WAVES

Radio burst signatures have been shown to correlate with SEP events (Cane et al., 2002; Gopalswamy et al., 2008). Type II radio bursts can indicate particle acceleration from CME-driven shocks, while Type III radio bursts are associated with accelerated electrons leaving the sun along open magnetic field lines. The *Wind* Radio and Plasma Wave Experiment (WAVES) instrument provides comprehensive coverage of radio and plasma wave phenomena in the frequency range of 20 kHz up to 13.825 MHz. We can identify radio burst signatures of accelerated particles using Radio Receiver Bands 1 and 2 (RAD1: 20–1,040 kHz; RAD2: 1.075–13.825 MHz). Only 20 channels of RAD1 contain observed data (the remaining channels are interpolated), and all 256 channels of RAD2 contain observed data; however, to reduce the data input, we use a total of 80 frequency channels from RAD1 and RAD2. In order to prevent CNNs from being driven by outlier pixels, we downsample the time axis to a cadence of 10 min, take the logarithm of the intensities, and clip them within the range (−1, 1). This process creates a single image of WAVES data 3-days before the flare onset in the form of a 432 time-frequency radio image. Figure 3 shows a 2D histogram of frequency observations from *Wind* between 20 Jan. 2005, and 23 Jan. 2005. At the beginning of the time interval, there is a clear type III radio burst.

2.2. In Situ Observations

In addition to remote observations, we provide a wide variety of near-Earth *in-situ* plasma measurements before, during, and after each observed flare event. We complement the Geostationary Operational Environmental Satellites (GOES) Flare Event List by measuring the integrated high-energy protons (≥5, ≥10, ≥30, ≥60, ≥100 MeV) using the GOES Energetic Particle Sensor (EPS; Sellers & Hanser, 1996), after 2017 the Solar and Galactic Proton Sensor on the Space Environment In Situ Suite. In addition to high-energy protons, we include the proton, alpha, and heavy ion fluxes in the suprathermal energy range (0.05–5.0 MeV). We measure the low-energy proton fluxes every 12 s with the Energetic Proton and Alpha Monitor (EPAM; Gold et al., 1998), and we measure the hourly-averaged alpha and heavy ion fluxes using the Ultra-Low Energy Ion Spectrometer (ULEIS; Mason et al., 1998) and the Solar Isotope Spectrometer (SIS; Stone et al., 1998). These instruments are part of a suite on board the Advanced Composition Explorer (ACE) (Stone et al., 1998), also stationed at L1. We also obtain the electron time series data at energies of 0.038–0.315 MeV from EPAM. For solar wind properties (i.e., velocity, temperature, density, magnetic field configuration), we utilize two instruments on board ACE; the local particle population data is from the Solar Wind Electron, Proton, and Alpha Monitor (McComas et al., 1998) with a 64-second time cadence, and local magnetic field vector properties from the Magnetic Field Experiment (MAG; Smith et al., 1998) with a 16-second time cadence.

2.2.1. Solar Wind

Solar wind conditions arrive at Earth sometimes delayed by several days after energetic solar events; however, in-situ measurements are critical to gauging the conditions of the solar atmosphere and interplanetary medium associated with these events. We include the first three solar wind moments (i.e., density, velocity, and temperature) and the vector components and magnitude of the interplanetary magnetic field. Disturbances within the already turbulent interplanetary medium can arise from CMEs and IP shocks, which are observed in the solar wind measurements taken at 1 au. These phenomena have the potential to modify particle acceleration and transportation during SEP events. Thus, the 24-hr averaged solar wind conditions provide insight into pre-flare solar activity, Solar Cycle (SC) phase, and expected route to 1 au. Figure 4, panels (e and f) show the solar wind speed, density, and temperature. The shaded region shows the time window used for parameter calculations. Panel (g) shows the magnetic field magnitude and vector quantities. The pre-flare conditions (shaded area), and the ICME signature is shown in the hatched, shaded area.

2.2.2. Suprathermal Population

Suprathermal particles are an extension of the solar wind tail at higher energies. These particles have energies ranging from a few keV to a few MeV. The origin of this population of particles is presently debatable (Cane et al., 2006; Mason & Sanderson, 1999; Tylka et al., 2005). They could be scattered remnants of large-scale transient events (Dayeh et al., 2009, 2017), or they could be accelerated stochastically in the solar atmosphere or in localized regions of high density in the solar wind (Fisk & Gloeckler, 2006). Nevertheless, they are nearly permanently present in the inner heliosphere, and they play an important role in post-flare SEP enhancement.

It is theorized that the highest energy particles in any given event already have suprathermal energies before the event. They are then re-accelerated by turbulence at the flare site or by propagating shocks (Desai et al., 2007; Fermi, 1949; Giacalone, 2005; Reames, 2017). Because of this phenomenon, we believe the ≥10 MeV time profiles strongly depend on the suprathermal seed population just before the flare onset (Cliver, 2006; Kahler, 2001). Therefore, we include the average pre-flare fluxes and abundance ratios of three of the most abundant energetic particle species (H, O, and Fe), which, to date, are not typically used in ML models. For all species, the lowest energy included is ∼1 MeV. The highest energy for H is 4.90 MeV, 8.50 MeV/nucleon for O, and 13.0 MeV/nucleon for Fe. The Fe/O Lo ratio energy is calculated at ∼1.13 eMeV/nucleon, and the Hi ratio energy is determined at ∼46 MeV/nucleon. Panel (d), Figure 4 shows the Oxygen (O) particle flux during a sample event, and the shaded area shows the sampling window used to calculate the parameters (i.e., Fe/O ratio) that are saved to the data set. The Hydrogen (H), Iron (Fe) abundances and ratios are done using the same sampling window.

2.2.3. X-Rays

Magnetic reconnection converts magnetic energy into particle kinetic energy. Particles accelerated down closed loops collide with the chromosphere and decelerate, producing large amounts of X-rays via Bremsstrahlung collisions (Galloway et al., 2010). X-rays can cause issues such as radio blackout or increased satellite drag (Xiong et al., 2014; Yasyukevich et al., 2018). X-rays are observed in the 0.05–0.4 and 0.1–0.8 nm wavelength bands by a primary and a secondary GOES satellite at a cadence of 60 s. We use the 0.1–0.8 nm wavelength band (also known as XL, the red line in Figure 4, panel a) to calculate the starting flux, peak flux (marked by the black X in Figure 4, panel a), rise time, fluence, and duration for each flare event. X-rays have been used in many SEP models including the Space Weather Prediction Center (SWPC), Air Force Research Laboratory, etc. (Balch, 2008; Kahler & Ling, 2015). Flare class prediction has been used in Support Vector Machines (Bobra & Couvidat, 2015; Choi et al., 2012; Nishizuka et al., 2018) and was used as a predictive output of one of the first attempts in neural networks by Fozzard (1989). There has been success using Long Short-Term Memory to predict X-ray time profiles (Li et al., 2020), and we find adding the X-ray time series beneficial to the ML model (shaded area of Figure 4, panel a).

2.2.4. High-Energy Protons and Electrons

Isotropic and permanently present high-energy protons in the heliosphere at energies above 10 MeV typically originate beyond the heliosphere as low-energy cosmic rays and their fluxes are too low to be considered a serious threat. However, strong X-ray flares on the Sun are associated with CMEs and shocks that efficiently accelerate protons to high energies and significantly increase radiation levels (Reames, 2013; Shea & Smart, 2012). This

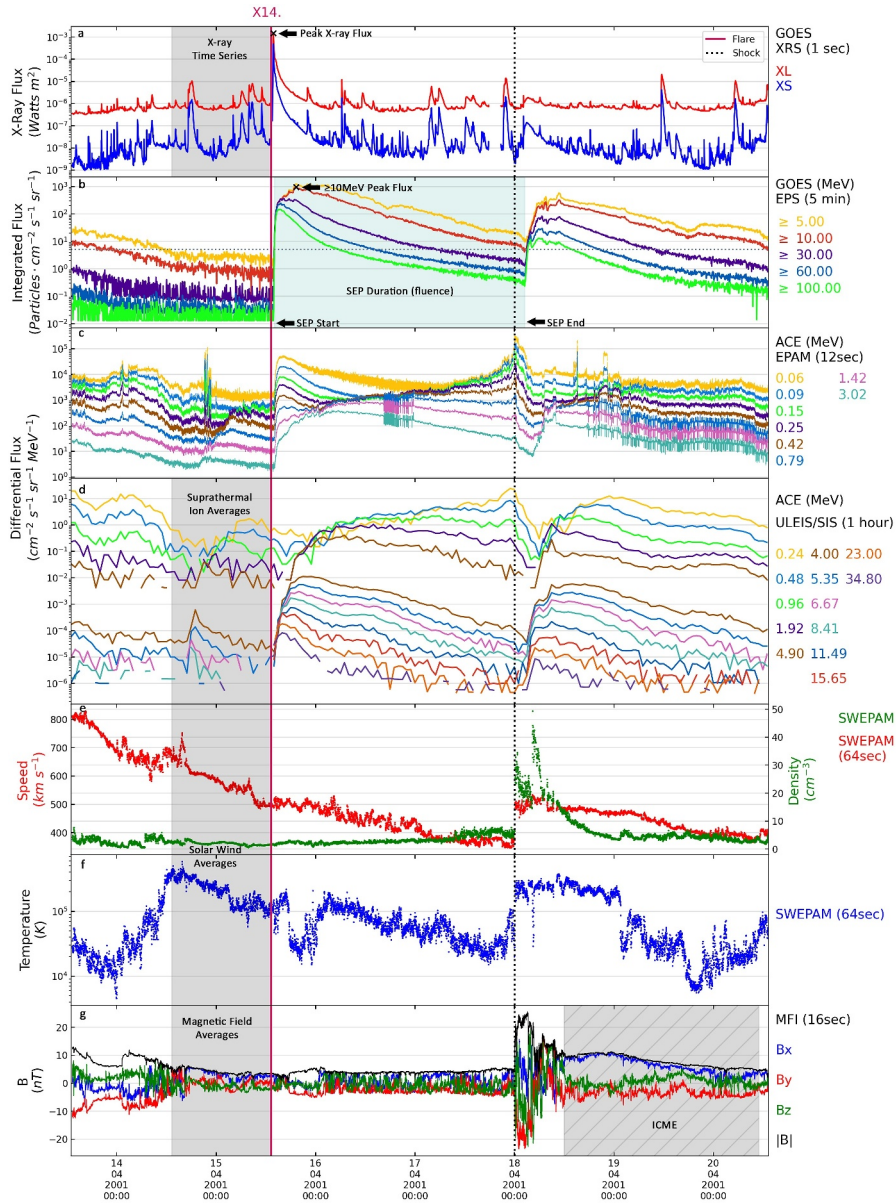

Figure 4. This plot is a sample of an SEP-positive event in the data set. Each panel pertains to a different instrument or observation. Panel (a) contains the X-ray flux, (b) shows the corrected integrated flux from GOES used to determine if a SEP event occurs and its subsequent properties, (c) at L1, we use proton flux from ACE to calculate properties during the shock, (d) including H, O, and Fe in low and high energies and ratios allows for observations of the suprathermal ion population before the event, (e, f) the pre-flare solar wind data is calculated in the pre-flare time window shown in the shaded area, (g) the magnitude and vectors of the interplanetary magnetic field are averaged and ICME (shaded, hatched area) properties are shown.

type of particulate radiation is of paramount concern and is consequently our primary measurement in determining event observations (see Section 4: SEP Event Detection). High energy electrons, though not energetic enough to seriously threaten human health, can cause spacecraft charging and damage electronics. Because of a significant speed difference, these relativistic electrons always arrive before the protons and are a strong predictor of the subsequent proton enhancement (Posner, 2007). Posner also notes that a SEP event has not been reported where the protons arrive at 1 au before the electrons, and 100 MeV ions arrive at a minimum of 13 min after the electrons. We, therefore, include electron measurements taken at L1 on ACE EPAM in the form of a time series at four energies (0.0535–0.2325 MeV) from the onset of the flare until the end of the flare event.

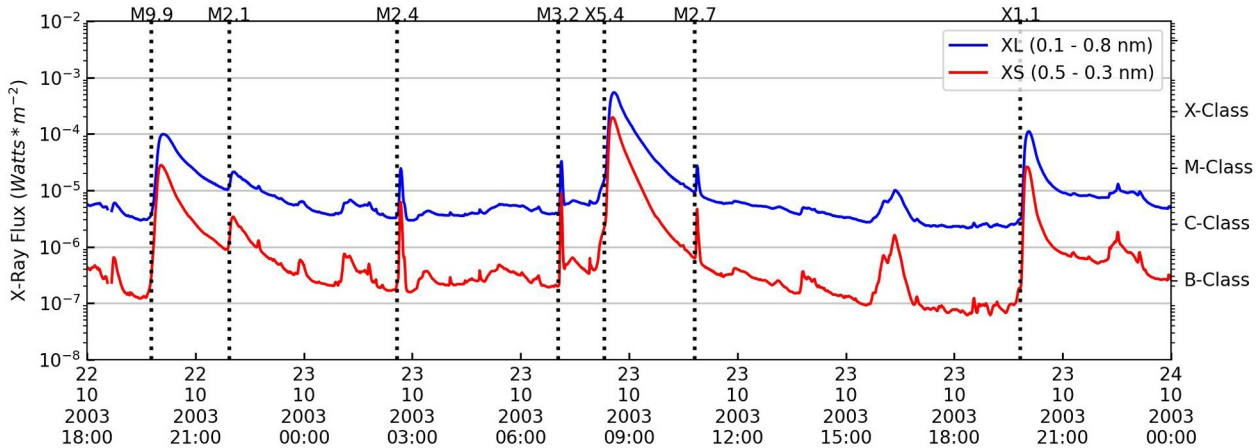

Figure 5. Sample X-ray time series from GOES flare event list with flare onset time marked by the dashed black vertical line and flare class identifier.

3. Event Lists

The National Oceanic and Atmospheric Administration (NOAA) began operating the GOES satellites with onboard X-ray and energetic particle detectors in 1975. Since then, all detected SFs have been cataloged by NOAA and are publicly available through the Heliophysics Events Knowledgebase (HEK). This list contains the GOES flare classification (e.g., X, M, C, B, A) and the flare properties such as start time, peak, and end time. When available, the list provides the flare location (latitude, longitude) and the corresponding AR (AR) number; flare events that happen beyond the solar limbs may not contain location or AR information. We obtain this list using SunPy's pre-built query of the GOES flare event list through the HEK (Barnes, Bobra, et al., 2020; Barnes, Cheung, et al., 2020). As of this publishing, the re-calibration of the GOES SF class data detailed in Cliver et al. (2022) has not been incorporated into our data set. Re-calibrated science-quality data from the National Centers for Environmental Information has been published to their website for GOES 13–15 but not for any other GOES spacecraft and is not yet ingested by the HEK database. Figure 5 shows a sample time series of the long (1.0–8.0°A) and short (0.5–3.0°A) X-ray bands observed by GOES with flares identified from the list. An explanation of the GOES event list flare detection algorithm is detailed in Ryan et al. (2016).

Interplanetary shock data is integrated from the Database of Heliospheric Shockwaves maintained at the University of Helsinki (Kilpua et al., 2015), which identifies over 2,600 shocks detected at ACE, WIND, and STEREO from 1975 to the present. The bolder, pink line expanded to all panels in Figure 4 marks an IP shock arrival during a sample event from the data set. CME parameters compiled since 1996 by Cane and Richardson (2003) are also ingested into the data set, and an example CME event is highlighted in the hatched shaded area in Figure 4, panel (g). For events after 2010, we can incorporate the Community Coordinated Modeling Center (CCMC) Space Weather Database of Notifications, Knowledge, Information (DONKI) web service to determine and verify linked events.

4. SEP Event Detection

We start the SEP event identification process by identifying a single flare in the NOAA/GOES flare event list discussed in Section 3. For each flare, we look for an enhancement in the ≥10 MeV corrected integrated flux within 6 hr of the flare onset. If an event were to originate from the Eastern limb, we estimate that it would take no longer than 6 hr for the ≥10 MeV SEPs to be observed. This estimate is based on an average time calculation from parent flare to SEP that is supported by other published SEP lists (e.g., Papaioannou et al., 2016; Rotti et al., 2022). If the flux of the proton enhancement exceeds 5 pfu (1 pfu = 1 proton cm^{-2} s^{-1} sr^{-1}) for 15 min (three consecutive data points) and the maximum pre-flare background flux (maximum flux value in the ≥10 MeV over a 3-hr time window before flare onset shown in the red highlight of Figure 5), then the event is considered a SEP-positive event. Checking the ≥10 MeV flux in the previous 3 hr is important for intervals of high activity; verifying the particle flux has become enhanced above the previous background, even when it is already over the 5 pfu threshold, is an important aspect of identifying new particle injection. If the flux has

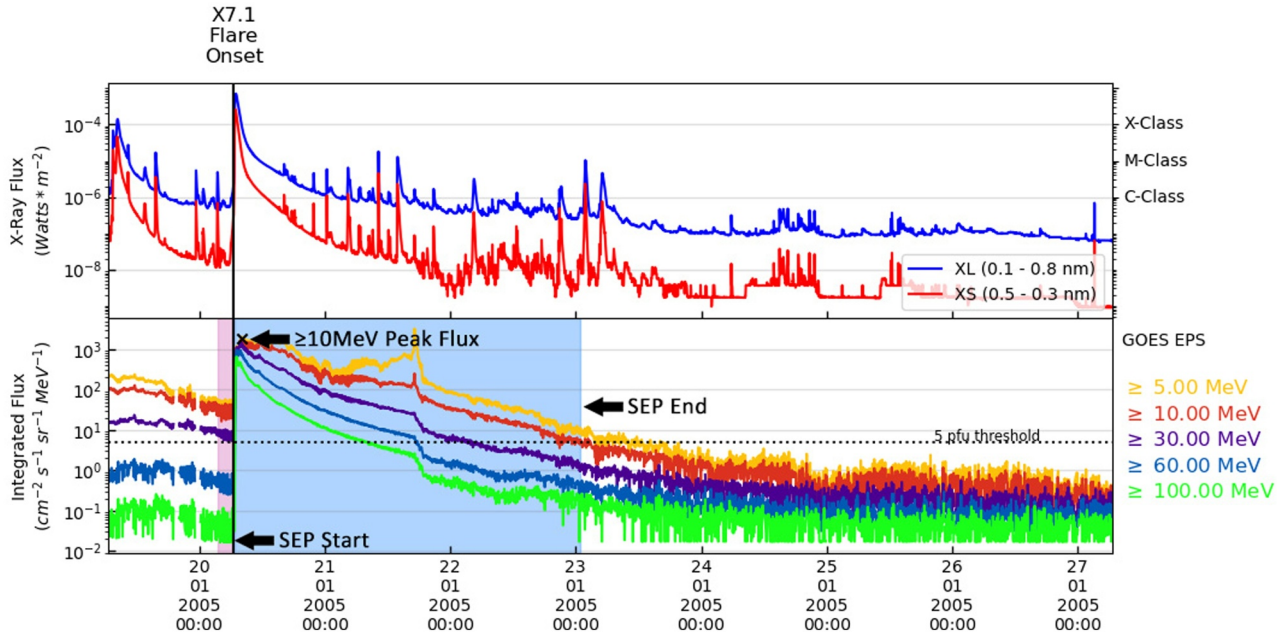

Figure 6. GOES X-ray and proton time series during a flare event marked as SEP-positive. The vertical black line marks the X7.1 flare onset, and the blue shaded area highlights the associated SEP event duration, the time frame the fluence is calculated over. The black X marks the peak flux value in the ≥10 MeV energy, and the 5 pfu threshold is marked by the horizontal dotted black line. The pink shaded area shows the 3-hr time window used to calculate the pre-flare flux values. To be positive, the SEP enhancement must also be larger than the max flux value pre-flare.

exceeded the threshold but is not greater than the flux in the previous 3-hr window, the event is marked as SEP-negative. We choose a conservative value of 5 pfu as the threshold energy instead of the typical 10 pfu, given by the NOAA SWPC definition for a proton event, because we are aiming for a conservative flux value that optimizes the number of positive events while still lying well above the instrument's background signal (Rodriguez et al., 2017). Figure 6 shows the profile of the flare and the SEP event for an event marked as SEP-positive.

In the event of multiple flares in a short time frame, the proton enhancement is attributed to the flare with the largest peak X-ray flux (Kahler, 1982). In the case of multiple flares (e.g., the event seen in 6, we set a flag in the data set (column "other flare"=1) that indicates this SEP event occurs in a time interval where other flares are present within a given time frame and may be contributing factors. Any flare events where the peak X-ray flux data is unavailable are removed from the data set.

For SEP-positive events, the start and stop time of the SEP are determined by the time the ≥10 MeV flux rises above the threshold for three consecutive points (15 min) and then falls below the threshold for three consecutive points (15 min). We then use the SEP start and stop time to calculate the duration of the SEP event and the time from flare onset to SEP onset. The proton flux is considered a SEP-negative event if it does not rise above the 5 pfu threshold. For SEP-negative events (Figure 7), parameters such as the start and stop time are filled with −9,999. The peak flux and fluence at five different energy ranges (≥5, ≥10, ≥30, ≥60, ≥100 MeV) are calculated for SEP-positive events over the duration of the event; for SEP-negative events, these values are calculated over a pre-determined time window of 6 hr (shown in the green highlight area of Figure 7. Note: the peak flux calculation may include the ESP enhancement due to associated IP shocks such as the one shown in Figure 6.

5. Data Validation and Limitations

Visualizing the SEP-positive and SEP-negative events allows us to compare our findings with those of previous studies, further validating our classification process. Of the 131 X-class flares, 41 are SEP-positive (31.3%). 71 out of 1649 M-class flares are SEP-positive (4.3%), and 140 out of 16014 C-class flares are SEP-positive (0.87%). Note the SEP-positive percentage rate increases with increasing flare class, as shown in Figure 8. This trend is consistent with the known correlations between flare strength and SEP enhancement (Kahler & Ling, 2018).

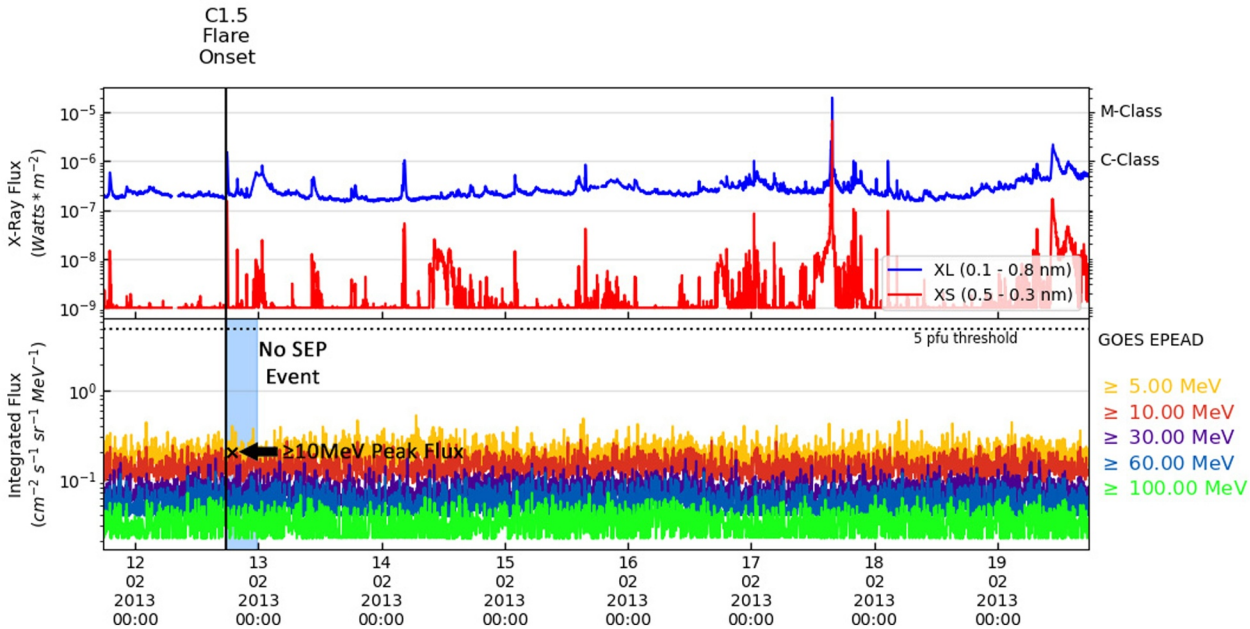

Figure 7. GOES X-ray and proton time series during a flare event marked as SEP-negative. The vertical black line marks the onset of the C1.5 flare. The ≥10 MeV flux does not exceed the 5 pfu threshold marked by the dotted black line during the preset time frame. Included values are calculated in the 6-hr window marked by the blue shaded area.

The distribution of SEP-positive events across flare classes exhibits a similar trend to those reported by NOAA SWPC within the same time frame (refer to the SPE.txt file available at ftp://ftp.swpc.noaa.gov/pub/indices/SPE.txt). In the SWPC list, C-class flares account for 13% of SEP-positive events, M-class flares for 51%, and X-class flares for 25%. Similarly, Papaioannou et al. (2016) identified 174 SEP-positive events spanning 1997–2013, with 17% from C-class flares, 43% from M-class flares, and 28% from X-class flares. The MEMPSEP data set includes a total of 252 SEP-positive events, with 50% attributed to C-class flares, 28% to M-class flares, and 16% to X-class flares. Notably, the proportion of SEP-positive events appears higher in the MEMPSEP data set for C-class flares; however, this is consistent with expectations given the conservative threshold of 5 p.f.u. employed in MEMPSEP data processing.

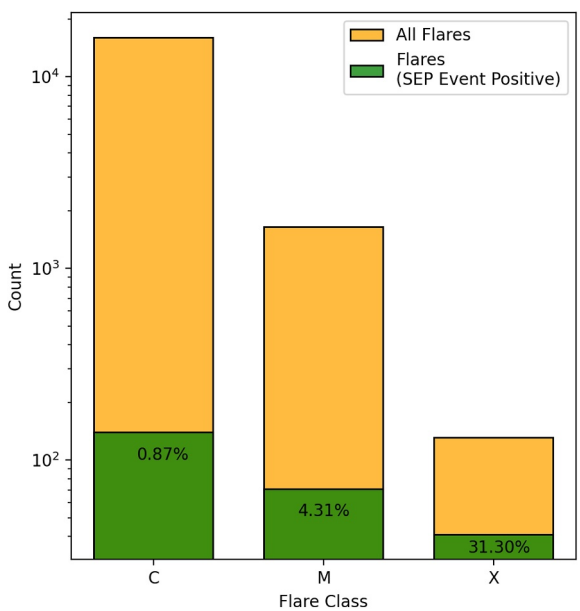

Figure 8. Class distribution of all GOES solar flare events (orange) by flare class from 1998 to 2013. The subset of SEP-positive flare events is shown in green, and their percentage with respect to all flares is noted in the respective bar. While the total number of SEP-positive flare events decreases with increasing flare class, the percentage of SEP-positive flare events with respect to the total number of flares in each class increases.

The association between sunspot number and SEP events has been well studied (e.g., Barnard & Lockwood, 2011; Birch & Bromage, 2022; Marroquin et al., 2023). Our data set shows a similar correlation between the number of SEP-positive events and the sunspot numbers reported by NOAA SWPC during SC 23 and SC 24. When considering larger SEP events (≥10 MeV particle flux exceeds the 10 pfu threshold), the pattern continues as shown in Figure 9. These findings reinforce the influence of solar activity, indicated by sunspot numbers, on the occurrence and magnitude of SEP events.

Flares originating from ARs west of the solar disk's central meridian are more likely to produce impactful SEPs near the Earth owing to Earth's magnetic connectivity to the CME-driven shock in or just beyond the solar corona (i.e., Ding et al., 2023; Kahler & Ling, 2018; Yashiro & Gopalswamy, 2008). Figure 10 illustrates the flare location on the solar disk compared with the logarithmic intensity of the ≥10 MeV particle energies using our data set.

Building accurate and reliable ML models hinges on the quality of the data fed into them. We address this challenge by rigorously validating our data set, ensuring it's free from instrument biases, errors, and inconsistencies.

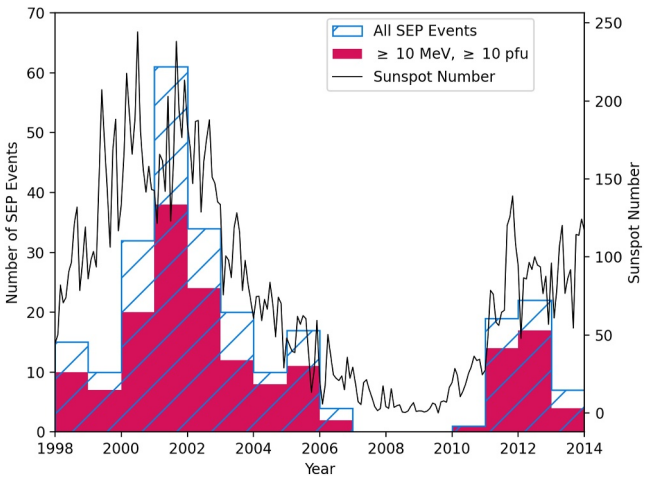

Figure 9. Histogram of the annual number of all SEP-Positive events where the ≥10 MeV protons cross the 5 pfu threshold (blue hashed bins) and SEP-positive events where the ≥10 MeV protons cross the 10 pfu threshold (filled pink bins). Monthly sunspot numbers are over-plotted in black. We find that the number of SEP-positive events and the subset of strong SEP-positive events follow the sunspot cycle.

For GOES data, the first step is verifying primary and secondary GOES satellites. Using the dates found on the NOAA documentation website (https://ngdc.noaa.gov/stp/satellite/goes/documentation.html), we create a file specifying which spacecraft is the primary or secondary for each instrument. Preference is given to data from the primary spacecraft. In the future, the data gathering routine will use the NOAA JSON data stream (https://services.swpc.noaa.gov/json/goes/instrument-sources.json) to determine the primary/secondary spacecraft. For GOES EPEAD, we pull the orientation flag file, which allows us to determine the correct variable name to use and the proper fill and max flux values. The max flux values allow us to check for bad data. Second, we identify and fill non-valid data for each instrument and set these values to −9,999. In cases where the missing data is used for calculating a vital parameter (e.g., the peak X-ray flux), the event is removed from the data set.

It is well known that the GOES particle sensors have high background levels due to the use of passive shielding (i.e., Hu & Semones, 2022; Kress et al., 2021; Rodriguez et al., 2017). As a result, particles can pass into the detectors from various locations and not just via the collimator and be counted. During SEP events, the low-energy passbands detect particles at the same time as the high-energy passbands, even though this is physically impossible. We do not subtract the background from the particle data. Instead, this data set uses the corrected integral proton data from GOES described in Rodriguez et al. (2017) when determining the SEP event properties such as onset time, peak flux, etc.

Data reduction techniques used in imaging, such as under-sampling, could impact forecasting in that small-scale transients might remain undetected. Using nearest-neighbor interpolation allows for straightforward identification of interpolated data based on the count of unique images in the stack. Therefore, we add a flag to any imaging data stacks where this interpolation technique has been used.

The in-situ data is labeled according to each process it belongs to (flare, pre-flare, SEP, etc.), saved, and made available in formats easily readable by ML algorithms. In situations where there is insufficient data to determine a property, such as the Fe/O ratio for a particular event, a fill value of −9,999 is assigned. However, it's important to note that fill values can potentially allow certain machine learning models to produce inaccurate results. Thus, the user may choose to exclude an event if a property is assigned a fill value. This will help ensure the accuracy and reliability of the data. Appendix A contains a table with the name and description of the data columns for the in-situ data that is available on Zenodo via the .csv file.

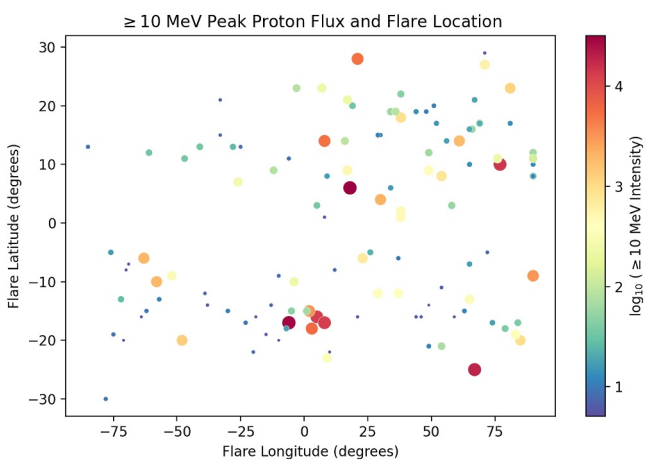

Figure 10. The solar disk location of each SEP-positive flare. The color and size indicate the logarithm of the ≥10 MeV proton intensity, indicated by the color bar. SEP-positive events tend to occur when the flare originates west of the central meridian, and the average intensity of the SEP-positive events is greater in the western hemisphere compared with the eastern hemisphere.

We recognize that different approaches can be taken for certain selections, like determining proton threshold values and associating parent flares. In a future release we will include a user-defined script, accessible via the Zenodo repository, that enables customization of these choices during data set generation. Using this script, the user can:

- Adjust the GOES proton threshold to a value of their choosing, that is, ≥10 MeV at 10 pfu.
- The averaging window for solar wind properties can be modified from 24 hr to user-defined time window.
- Refine flare association by checking for the flare with the best Earth connectivity within a specified time window, provided all flares in that window have valid location data.
- The option to remove events where the SEP end time is the same (i.e., events that might occur at the tail of a previous event) or to connect the event to a parent flare based on criteria such as first flare or largest flare.

6. Future Work

Events from SC 24 and 25, spanning the years 2013–2022, are being added to the next iteration of the data set. Parameters from the lists described in Section 3 will be carefully linked to their parent events.

The inclusion of the associated CMEs, IP Shocks, and ESP events will allow for a more complete data set as well as the ability to set flags to notify the user of the data set about possible issues with the data values (e.g., if an associated ESP causes a proton enhancement in the GOES data, a flag to notify the user that the peak flux during that event includes the ESP would be set).

The addition of meaningful parameters will continue, and their data description will be added as the data set matures. Currently, we are incorporating post-flare EPAM proton data in addition to EPAM electron data. Important observables such as the strength and direction of the magnetic field, magnetic field fluctuations, and turbulence, which can play a role in particle transport, are being considered. With the anticipated launches of the Interstellar Mapping and Acceleration Probe (IMAP; McComas et al., 2018) and the Space Weather Follow-On—Lagrange 1 (SWFO-L1), high-resolution near real-time measurements and unprecedented insights into the dynamics of the solar wind and interplanetary medium will become available, offering opportunities to expand and enrich the ML data set established in this study and its applications.

7. Conclusions

In this study, we have created a meticulously selected data set of SEP-positive and SEP-negative events that took place between 1998 and 2013. We have included in-situ and remote observations associated with these events for the purpose of using them in machine learning models. To ensure the accuracy of our data set, we have cross-checked it with other published lists and verified known trends of SEP events. From 1998 to 2013, over 19,000 flare events were classified as C-class or higher. Among these, 252 flares were associated with a proton event where the ≥10 MeV integrated flux crosses a threshold of 5 pfu.

We used coding procedures that utilized established parameters to categorize events as SEP-positive or SEP-negative. This approach minimizes human bias in identifying SEP events and ensures consistent and scalable analysis. For each event, a set of in-situ data is obtained, and relevant properties are derived. Additionally, imaging stacks and time series are generated for each event. The data is packaged and will be fully available via the public Zenodo link. The data set, in this current form, served as input for a machine-learning ensemble model, MEMPSEP, used to predict SEP onset times and other properties such as peak flux.

Our data set aims to minimize the time researchers spend on data curation, allowing them to focus on the scientific aspects of SEPs. We provide a comprehensive list of parameters that correspond to each time step of the SEP events process, including both in-situ and remote imaging data. Additionally, we plan to collect data from upcoming missions to ensure that a wide range of information is available for machine learning models.

Appendix A

Column name	Description	Time window
FlrOnset	Date and time of the flare start	
Rise Time	Time from flare start to flare peak (in fractional hours)	
Starting Flux	X-ray flux value at start of flare	
Peak Flux	X-ray flux value at peak of flare	
Duration	Time from flare onset to flare end	
Fluence	Fluence of the flare from start to end	
SW Temp	Temperature of the solar wind (K)	24 hr average
SW Velocity	Velocity of the solar wind (km/s)	24 hr average
SW Density	Density of the solar wind (cm^{-3})	24 hr average
IMF B	Magnitude of the magnetic field at 1 au	24 hr average
IMF Bx	X-component of the magnetic field at 1 au	24 hr average

Continued

Column name	Description	Time window
IMF By	Y-component of the magnetic field at 1 au	24 hr average
IMF Bz	Z-component of the magnetic field at 1 au	24 hr average
Fe/O Lo	Iron to Oxygen ratio (~0.11 MeV)	24 hr average
Fe/O Hi	Iron to Oxy gen ratio (~46 MeV)	24 hr average
H Lo	Hydrogen (~1 MeV)	24 hr average
H Hi	Hydrogen (4.90 MeV)	24 hr average
O Lo	Oxygen (~1 MeV)	24 hr average
O Hi	Oxygen (8.50 MeV)	24 hr average
Fe Lo	Iron (~1 MeV)	24 hr average
Fe Hi	Iron (13.0 MeV)	24 hr average
Fe/O 1 MeV	Iron to Oxygen ratio (~1 MeV)	24 hr average
event date	Date of the flare	
start time	Start time of flare	
peak time	Peak time of flare	
end time	End time of flare	
Xrsclass	Major class of flare	
goes _location	Longitude and Latitude of flare (if available)	
noaa active region	NOAA SWPCs Active Region Number (if available)	
Flrmaxtime	Time of flare peak	
Flrendtime	Time of flare end	
goes _major class	GOES Flare Classification	
SEP-Server Association	if = "ProtonFlare" event is marked as SEP-positive from PSEP list Papaioannou et al. (2016)	
FlrLon	Flare longitude (parsed from goes location, if available)	
FlrLat	Flare latitude (passed from goes location, if available)	
event type	SEP-positive = 1, SEP-negative = 0	
other flare	if = 1 other flares present within 2-hour time window sep	
peak 1	Peak flux of SEP event (≥5 MeV)	
sep fluence _1	Fluence of SEP event (≥5 MeV)	
sep peak 2	Peak flux of SEP event (≥10 MeV)	
sep fluence _2	Fluence of SEP event (≥10 MeV)	
sep peak 3	Peak flux of SEP event (≥30 MeV)	
sep fluence _3	Fluence of SEP event (≥30 MeV)	
sep peak 4	Peak flux of SEP event (≥50 MeV)	
sep fluence _4	Fluence of SEP event (≥50 MeV)	
sep peak 5	Peak flux of SEP event (≥100 MeV)	
sep fluence _5	Fluence of SEP event (≥100 MeV)	
sep start	Start of SEP event (onset)	
sep end	End of SEP event	
sep rise	Time from start of SEP to peak (fractional hours)	
sep duration	Time from start of SEP to end (fractional hours)	

Data Availability Statement

The "X-ray Flare" data set and the GOES X-ray sensor data were prepared and made available through the NOAA National Geophysical Data Center (NGDC)—(https://www.ngdc.noaa.gov/stp/satellite/goes/). We acknowledge the use of NOAA's 1–8°A solar X-ray data and proton data from the NOAA GOES data archive. The MDI and AIA data is used courtesy of NASA/SDO and the AIA, EVE, and HMI science teams (https://sdo.gsfc.nasa.gov/data/aiahmi/). The SOHO/LASCO data used here are produced by a consortium of the Naval Research Laboratory (USA) MaxPlanck-Institut fuer Aeronomie (Germany), Laboratoire d'Astronomie (France), and the University of Birmingham (UK)—(https://www.cosmos.esa.int/web/soho/soho-science-archive). SOHO is an international cooperation project between ESA and NASA. This paper uses data from the Heliospheric Shock Database, generated and maintained at the University of Helsinki (http://ipshocks.fi/).

A .csv file containing the flare, solar wind, suprathermal, and SEP event data is currently available on Zenodo (https://zenodo.org/records/10044865) under a Creative Commons Attribution license at (Moreland & Dayeh, 2023). The data set also includes the imaging stacks, and time series data. In a future version a customization scripts will enable users to define variables and generate their own unique data sets. Access to these resources will be provided through the Zenodo repository.

Software: This data set used version 4.0 of SunPy open source software package (Barnes, Bobra, et al., 2020; Barnes, Cheung, et al., 2020), version 3.5.1 matplotlib (Hunter, 2007), version 1.21.2 NumPy (Harris et al., 2020), version 1.4.1 pandas (pandas development team, 2020), version 1.7.2 SciPy (Virtanen et al., 2020), version 0.11.2 seaborn (Waskom, 2021), and customized download scripts written by Samuel Hart.

Acknowledgments
This work is mainly supported by SWO2R Grant 80NSSC20K0290. Partial support for MAD, KM, and STH comes from SW2OR 80NSSC21K0027, NASA LWS Grants 80NSSC19K0079, 80NSSC21K1307, and 80NSSC20K1815.

References

Angryk, R. A., Martens, P. C., Aydin, B., Kempton, D., Mahajan, S. S., Basodi, S., et al. (2020). Multivariate time series dataset for space weather data analytics. *Scientific Data*, *7*(1), 227. https://doi.org/10.1038/s41597-020-0548-x

Balch, C. C. (2008). Updated verification of the space weather prediction center's solar energetic particle prediction model. *Space Weather*, *6*(1), S01001. https://doi.org/10.1029/2007SW000337

Barnard, L., & Lockwood, M. (2011). A survey of gradual solar energetic particle events. *Journal of Geophysical Research*, *116*(A5), A05103. https://doi.org/10.1029/2010JA016133

Barnes, W. T., Bobra, M. G., Christe, S. D., Freij, N., Hayes, L. A., Ireland, J., et al. (2020a). The SunPy project: Open source development and status of the version 1.0 core package. *The Astrophysical Journal*, *890*(1), 68. https://doi.org/10.3847/1538-4357/ab4f7a

Barnes, W. T., Cheung, M. C. M., Bobra, M. G., Boerner, P. F., Chintzoglou, G., Leonard, D., et al. (2020b). aiapy: A python package for analyzing solar EUV image data from AIA. *Journal of Open Source Software*, *5*(55), 2801. https://doi.org/10.21105/joss.02801

Birch, M., & Bromage, B. (2022). Sunspot numbers and proton events in solar cycles 19 to 24. *Journal of Atmospheric and Solar-Terrestrial Physics*, *236*, 105891. https://doi.org/10.1016/j.jastp.2022.105891

Bobra, M. G., & Couvidat, S. (2015). Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, *798*(2), 135. https://doi.org/10.1088/0004-637x/798/2/135

Bobra, M. G., Wright, P. J., Sun, X., & Turmon, M. J. (2021). SMARPs and SHARPs: Two solar cycles of active region data. *The Astrophysical Journal - Supplement Series*, *256*(2), 26. https://doi.org/10.3847/1538-4365/ac1f1d

Bougeret, J. L., Kaiser, M. L., Kellogg, P. J., Manning, R., Goetz, K., Monson, S. J., et al. (1995). WAVES: The radio and plasma wave investigation on the wind spacecraft. *Space Science Reviews*, *71*(1–4), 231–263. https://doi.org/10.1007/BF00751331

Brueckner, G., Howard, R., Koomen, M., Korendyke, C., Michels, D., Moses, J., et al. (1995). Visible light coronal imaging and spectroscopy. *Solar Physics*, *162*(1–2), 357–402. https://doi.org/10.1007/bf00733434

Cairns, I. H., Lobzin, V. V., Donea, A., Tingay, S. J., McCauley, P. I., Oberoi, D., et al. (2018). Low altitude solar magnetic reconnection, type III solar radio bursts, and x-ray emissions. *Scientific Reports*, *8*(1), 1612–1676. https://doi.org/10.1038/s41598-018-19195-3

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*(8), 1166–1207. https://doi.org/10.1029/2018SW002061

Cane, H. V., Erickson, W. C., & Prestage, N. P. (2002). Solar flares, type iii radio bursts, coronal mass ejections, and energetic particles. *Journal of Geophysical Research*, *107*(A10), 1315. https://doi.org/10.1029/2001JA000320

Cane, H. V., Mewaldt, R. A., Cohen, C. M. S., & von Rosenvinge, T. T. (2006). Role of flares and shocks in determining solar energetic particle abundances. *Journal of Geophysical Research*, *111*(A6), A06S90. https://doi.org/10.1029/2005JA011071

Cane, H. V., & Richardson, I. G. (2003). Interplanetary coronal mass ejections in the near earth solar wind during 1996–2002. *Journal of Geophysical Research*, *108*(A4), 1156. https://doi.org/10.1029/2002JA009817

Chancellor, J. C., Scott, G. B., & Sutton, J. P. (2014). Space radiation: The number one risk to astronaut health beyond low earth orbit. *Life*, *4*(3), 491–510. https://doi.org/10.3390/life4030491

Chatterjee, S., Dayeh, M., Mun˜oz-Jaramillo, A., Bain, H. M., Moreland, K., & Hart, S. (2024). Mempsep-I: Forecasting the probability of solar energetic particle event occurrence using a multivariate ensemble of convolutional neural networks. *Space Weather*. https://doi.org/10.1029/2023SW003568

Chatterjee, S., Mun˜oz-Jaramillo, A., Dayeh, M. A., Bain, H. M., & Moreland, K. (2023). Homogenizing SOHO/EIT and SDO/AIA 171 ˚A images: A deep-learning approach. *The Astrophysical Journal - Supplement Series*, *268*(1), 33. https://doi.org/10.3847/1538-4365/ace9d7

Choi, J.-H., Choi, C., Ko, B.-K., & Kim, P.-K. (2012). Detection of cross site scripting. Attack in wireless networks using n-Gram and SVM. *Mobile Information Systems*, *8*(3), 275–286. https://doi.org/10.3233/MIS-2012-0143

Cliver, E. (2006). The unusual relativistic solar proton events of 1979 august 21 and 1981 may 10. *The Astrophysical Journal*, *639*(2), 1206–1217. https://doi.org/10.1086/499765

Cliver, E., Dennis, B. R., Kiplinger, A. L., Kane, S. R., Neidig, D. F., Sheeley, J. N. R., & Koomen, M. J. (1986). Solar gradual hard x-ray bursts and associated phenomena. *The Astrophysical Journal*, *305*, 920. https://doi.org/10.1086/164306

Cliver, E. W., Schrijver, C. J., Shibata, K., & Usoskin, I. G. (2022). Extreme solar events. *Living Reviews in Solar Physics*, *19*(1), 2. https://doi.org/10.1007/s41116-022-00033-8

Crosby, N., Heynderickx, D., Jiggens, P., Aran, A., Sanahuja, B., Truscott, P., et al. (2015). SEPEM: A tool for statistical modeling the solar energetic particle environment. *Space Weather*, *13*(7), 406–426. https://doi.org/10.1002/2013SW001008

Dayeh, M. A., Chatterjee, S., Munoz-Jaramillo, A., Moreland, K., Bain, H. M., & Hart, S. (2024). Mempsep-II—Forecasting the properties of solar energetic particle events using a multivariate ensemble approach. *Space Weather*. https://doi.org/10.1029/2023SW003697

Dayeh, M. A., Desai, M. I., Dwyer, J. R., Rassoul, H. K., Mason, G. M., & Mazur, J. E. (2009). Composition and spectral properties of the 1 AU QuietTime suprathermal ion population during solar cycle 23. *The Astrophysical Journal*, *693*(2), 1588–1600. https://doi.org/10.1088/0004-637X/693/2/1588

Dayeh, M. A., Desai, M. I., Ebert, R. W., Elliott, H., Farahat, A., Kozarev, K., & Li, G. (2018). What causes the variability in the properties of energetic storm particle (ESP) events? *Journal of Physics: Conference Series*, *1100*(1), 012008. https://doi.org/10.1088/1742-6596/1100/1/012008

Dayeh, M. A., Desai, M. I., Mason, G. M., Ebert, R. W., & Farahat, A. (2017). Origin and properties of quiet-time 0.11–1.28 Mev Nucleon1 heavy-ion population near 1 au. *The Astrophysical Journal*, *835*(2), 155. https://doi.org/10.3847/1538-4357/835/2/155

Decker, R. B. (1981). The modulation of low-energy proton distributions by propagating interplanetary shock waves: A numerical simulation. *Journal of Geophysical Research*, *86*(A6), 4537–4554. https://doi.org/10.1029/JA086iA06p04537

Delaboudinière, J. P., Artzner, G. E., Brunaud, J., Gabriel, A. H., Hochedez, J. F., Millier, F., et al. (1995). EIT: Extreme-ultraviolet Imaging Telescope for the SOHO mission. *Solar Physics*, *162*(1-2), 291–312. https://doi.org/10.1007/BF00733432

Desai, M., & Giacalone, J. (2016). Large gradual solar energetic particle events. *Living Reviews in Solar Physics*, *13*(1), 3. https://doi.org/10.1007/s41116-016-0002-5

Desai, M. I., Mason, G. M., Dayeh, M. A., Ebert, R. W., McComas, D. J., Li, G., et al. (2016). Charge-to-mass dependence of heavy ion spectral breaks in large gradual solar energetic particle events. *Journal of Physics: Conference Series*, *767*(1), 012004. https://doi.org/10.1088/1742-6596/767/1/012004

Desai, M. I., Mason, G. M., Gold, R. E., Krimigis, S. M., Cohen, C. M. S., Mewaldt, R. A., et al. (2007). Evidence for a two-stage acceleration process in large solar energetic particle events. *Space Science Reviews*, *130*(1), 243–253. https://doi.org/10.1007/s11214-007-9219-x

Ding, Z., Li, G., Santa Fe Duen˜as, A., Ebert, R. W., Wijsen, N., & Poedts, S. (2023). The east-west asymmetry of particle intensity in energetic storm particle events. *Journal of Geophysical Research: Space Physics*, *128*(7), e2023JA031502. https://doi.org/10.1029/2023JA031502

Domingo, V., Fleck, B., & Poland, A. I. (1995). SOHO: The solar and heliospheric observatory. *Space Science Reviews*, *72*(1–2), 81–84. https://doi.org/10.1007/BF00768758

Fermi, E. (1949). On the origin of the cosmic radiation. *Physical Review*, *75*(8), 1169–1174. https://doi.org/10.1103/PhysRev.75.1169

Fisk, L. A., & Gloeckler, G. (2006). The common spectrum for accelerated ions in the quiet-time solar wind. *The Astrophysical Journal*, *640*(1), L79–L82. https://doi.org/10.1086/503293

Fozzard, R. (1989). Theonet: A connectionist expert system for solar flare forecasting (doctoral dissertation, university of Colorado, boulder, department of computer science). Retrieved from https://spl.cde.state.co.us/artemis/ucbserials/ucb51110internet/1989/ucb51110442internet.pdf

Galloway, R., Helander, P., MacKinnon, A., & Brown, J. (2010). Thermalisation and hard X-ray bremsstrahlung efficiency of self-interacting solar flare fast electrons. *Astronomy & Astrophysics*, *520*, A72. https://doi.org/10.1051/0004-6361/201014077

Galvez, R., Fouhey, D. F., Jin, M., Szenicer, A., Mun˜oz-Jaramillo, A., Cheung, M. C. M., et al. (2019). A machine-learning data set prepared from the NASA solar dynamics observatory mission. *The Astrophysical Journal Supplement Series*, *242*(1), 7. https://doi.org/10.3847/1538-4365/ab1005

Giacalone, J. (2005). 5). Particle acceleration at shocks moving through an irregular magnetic field. *The Astrophysical Journal*, *624*(2), 765–772. https://doi.org/10.1086/429265

Giacalone, J., & Neugebauer, M. (2008). The energy spectrum of energetic particles downstream of turbulent collisionless shocks. *The Astrophysical Journal*, *673*(1), 629–636. https://doi.org/10.1086/524008

Gold, R. E., Krimigis, S. M., Hawkins, S. E., Haggerty, D. K., Lohr, D. A., Fiore, E., et al. (1998). Electron, proton, and alpha monitor on the advanced composition explorer spacecraft. *Space Science Reviews*, *86*(1–4), 541–562. https://doi.org/10.1007/978-94-011-4762-019

Gopalswamy, N., Xie, H., Akiyama, S., M¨akel¨a, P. A., & Yashiro, S. (2014). Major solar eruptions and high-energy particle events during solar cycle 24. *Earth Planets and Space*, *66*(1), 104. https://doi.org/10.1186/1880-5981-66-104

Gopalswamy, N., Yashiro, S., Akiyama, S., M¨akel¨a, P., Xie, H., Kaiser, M. L., et al. (2008). Coronal mass ejections, type ii radio bursts, and solar energetic particle events in the SOHO era. *Annales Geophysicae*, *26*(10), 3033–3047. https://doi.org/10.5194/angeo-26-3033-2008

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Horne, R. B., Glauert, S. A., Meredith, N. P., Boscher, D., Maget, V., Heynderickx, D., & Pitchford, D. (2013). Space weather impacts on satellites and forecasting the earth's electron radiation belts with spacecast. *Space Weather*, *11*(4), 169–186. https://doi.org/10.1002/swe.20023

Hu, S., & Semones, E. (2022). Calibration of the goes 6–16 high-energy proton detectors based on modelling of ground level enhancement energy spectra. *Journal of Space Weather and Space Climate*, *12*, 5. https://doi.org/10.1051/swsc/2022003

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kahler, S. W. (1982). The role of the big flare syndrome in correlations of solar energetic proton fluxes and associated microwave burst parameters. *Journal of Geophysical Research*, *87*(A5), 3439–3448. https://doi.org/10.1029/JA087iA05p03439

Kahler, S. W. (2001). The correlation between solar energetic particle peak intensities and speeds of coronal mass ejections: Effects of ambient particle intensities and energy spectra. *Journal of Geophysical Research*, *106*(A10), 20947–20955. https://doi.org/10.1029/2000JA002231

Kahler, S. W., & Ling, A. (2015). Dynamic SEP event probability forecasts. *Space Weather*, *13*(10), 665–675. https://doi.org/10.1002/2015SW001222

Kahler, S. W., & Ling, A. G. (2018). Forecasting solar energetic particle (SEP) events with flare X-ray peak ratios. *Journal of Space Weather and Space Climate*, *8*, 1–10. https://doi.org/10.1051/swsc/2018033

Kahler, S. W., White, S. M., & Ling, A. G. (2017). Forecasting E ¿ 50-MeV proton events with the proton prediction system (PPS). *Journal of space weather and space climate*, *7*, A27. https://doi.org/10.1051/swsc/2017025

Kilpua, E. K., Lumme, E., Andreeova, K., Isavnin, A., & Koskinen, H. E. (2015). Properties and drivers of fast interplanetary shocks near the orbit of the Earth (1995-2013). *Journal of Geophysical Research: Space Physics*, *120*(6), 4112–4125. https://doi.org/10.1002/2015JA021138

Kress, B. T., Rodriguez, J. V., Boudouridis, A., Onsager, T. G., Dichter, B. K., Galica, G. E., & Tsui, S. (2021). Observations from NOAA'S newest solar proton sensor. *Space Weather*, *19*(12), e2021SW002750. https://doi.org/10.1029/2021SW002750

Kutsenko, A. S., Abramenko, V. I., & Pevtsov, A. A. (2019). Extended statistical analysis of emerging solar active regions. *Monthly Notices of the Royal Astronomical Society*, *484*(3), 4393–4400. https://doi.org/10.1093/mnras/stz308

Kutsenko, A. S., Abramenko, V. I., & Plotnikov, A. A. (2021). A statistical study of magnetic flux emergence in solar active regions prior to strongest flares.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, M. A. (1983). Coupled hydromagnetic wave excitation and ion acceleration at interplanetary traveling shocks. *Journal of Geophysical Research*, *88*(A8), 6109–6119. https://doi.org/10.1029/JA088iA08p06109

Lemen, J. R., Title, A. M., Akin, D. J., Boerner, P. F., Chou, C., Drake, J. F., et al. (2012). The atmospheric imaging assembly (AIA) on the solar dynamics observatory (SDO). *Solar Physics*, *275*(1–2), 17–40. https://doi.org/10.1007/s11207-011-9776-8

Li, X., Zheng, Y., Wang, X., & Wang, L. (2020). Predicting solar flares using a novel deep convolutional neural network. *The Astrophysical Journal*, *891*(1), 10. https://doi.org/10.3847/1538-4357/ab6d04

Mäkelä, P., Gopalswamy, N., Akiyama, S., Xie, H., & Yashiro, S. (2011). Energetic storm particle events in coronal mass ejection-driven shocks. *Journal of Geophysical Research*, *116*(8), 1–12. https://doi.org/10.1029/2011JA016683

Marroquin, R. D., Sadykov, V., Kosovichev, A., Kitiashvili, I. N., Oria, V., Nita, G. M., et al. (2023). Statistical study of the correlation between solar energetic particles and properties of active regions. *The Astrophysical Journal*, *952*(2), 97. https://doi.org/10.3847/1538-4357/acdb65

Mason, G. M., Gold, R. E., Krimigis, S. M., Mazur, J. E., Andrews, G. B., Daley, K. A., et al. (1998). The ultra-low-energy isotope spectrometer (ULEIS) for the ace spacecraft. *Space Science Reviews*, *86*(1–4), 409–448. https://doi.org/10.1007/978-94-011-4762-016

Mason, G. M., & Sanderson, T. R. (1999). CIR associated energetic particles in the inner and middle heliosphere. *Space Science Reviews*, *89*(1), 77–90. https://doi.org/10.1023/A:1005216516443

Maurer, R. H., Fretz, K., Angert, M. P., Bort, D. L., Goldsten, J. O., Ottman, G., et al. (2017). Radiation-induced single-event effects on the Van Allen probes spacecraft. *IEEE Transactions on Nuclear Science*, *64*(11), 2782–2793. https://doi.org/10.1109/TNS.2017.2754878

McComas, D. J., Bame, S. J., Barker, P., Feldman, W. C., Phillips, J. L., Riley, P., & Griffee, J. W. (1998). Solar wind electron proton alpha monitor (SWEPAM) for the advanced composition explorer. *Space Science Reviews*, *86*(1–4), 563–612. https://doi.org/10.1007/978-94-011-4762-020

McComas, D. J., Christian, E. R., Schwadron, N. A., Fox, N., Westlake, J., Allegrini, F., et al. (2018). Interstellar mapping and acceleration Probe (IMAP): A new NASA mission. *Space Science Reviews*, *214*(8), 116. https://doi.org/10.1007/s11214-018-0550-1

Moreland, K., & Dayeh, M. (2023). MEMSEP dataset (1998-2013) (1.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.10044865

Moreland, K., Dayeh, M. A., Li, G., Farahat, A., Ebert, R. W., & Desai, M. I. (2023). Variability of interplanetary shock and associated energetic particle properties as a function of the time window around the shock. *The Astrophysical Journal*, *956*(2), 107. https://doi.org/10.3847/1538-4357/acec6c

Mostafavi, P., Zank, G. P., & Webb, G. M. (2018). The mediation of collisionless oblique magnetized shocks by energetic particles. *The Astrophysical Journal*, *868*(2), 120. https://doi.org/10.3847/1538-4357/aaeb91

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. (2018). Deep flare net (DeFN) model for solar flare prediction. *The Astrophysical Journal*, *858*(2), 113. https://doi.org/10.3847/1538-4357/aab9a7

Onorato, G., Di Schiavi, E., & Di Cunto, F. (2020). Understanding the effects of deep space radiation on nervous system: The role of genetically tractable experimental models. *Frontiers in Physics*, *8*(October), 1–11. https://doi.org/10.3389/fphy.2020.00362

pandas development team, T. (2020). pandas-dev/pandas: Pandas. *Zenodo*. https://doi.org/10.5281/zenodo.3509134

Papaioannou, A., Sandberg, I., Anastasiadis, A., Kouloumvakos, A., Georgoulis, M. K., Tziotziou, K., et al. (2016). Solar flares, coronal mass ejections and solar energetic particle event characteristics. *Journal of Space Weather and Space Climate*, *6*, A42. https://doi.org/10.1051/swsc/2016035

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. (2012). The solar dynamics observatory (SDO). *Solar Physics*, *275*(1–2), 3–15. https://doi.org/10.1007/s11207-011-9841-3

Posner, A. (2007). Up to 1-hour forecasting of radiation hazards from solar energetic ion events with relativistic electrons. *Space Weather*, *5*(5). https://doi.org/10.1029/2006SW000268

Reames, D. V. (2001). Seps: Space weather hazard in interplanetary space. In *Space weather* (pp. 101–107). American Geophysical Union (AGU). https://doi.org/10.1029/GM125p0101

Reames, D. V. (2013). The two sources of solar energetic particles. *Space Science Reviews*, *175*(1–4), 53–92. https://doi.org/10.1007/s11214-013-9958-9

Reames, D. V. (2017). The abundance of helium in the source plasma of solar energetic particles. *Solar Physics*, *292*(11), 1–20. https://doi.org/10.1007/s11207-017-1173-5

Rodriguez, J. V., Sandberg, I., Mewaldt, R. A., Daglis, I. A., & Jiggens, P. (2017). Validation of the effect of cross-calibrated goes solar proton effective energies on derived integral fluxes by comparison with stereo observations. *Space Weather*, *15*(2), 290–309. https://doi.org/10.1002/2016SW001533

Rotti, S., Aydin, B., Georgoulis, M. K., & Martens, P. C. (2022). Integrated geostationary solar energetic particle events catalog: GSEP. *The Astrophysical Journal Supplement Series*, *262*(1), 29. https://doi.org/10.3847/1538-4365/ac87ac

Ryan, D. F., Dominique, M., Seaton, D., Stegen, K., & White, A. (2016). Effects of flare definitions on the statistics of derived flare distributions. *Astronomy and Astrophysics*, *592*, A133. https://doi.org/10.1051/0004-6361/201628130

Scherrer, P. H., Bogart, R. S., Bush, R. I., Hoeksema, J. T., Kosovichev, A. G., Schou, J., et al. (1995). The solar oscillations investigation—Michelson Doppler imager. *Solar Physics*, *162*(1), 129–188. https://doi.org/10.1007/BF00733429

Schou, J., Scherrer, P. H., Akin, D. J., Allard, B. A., Miles, J. W., Rairden, R., et al. (2012). Design and ground calibration of the helioseismic and magnetic imager (HMI) instrument on the solar dynamics observatory (SDO) NASA Center for AeroSpace Information (CASI).Journal Articles. *Solar Physics*, *275*(1–2), 229–259. (Copyright - Copyright NASA/Langley Research Center Jan 1, 2012; Last updated - 2018-12-13). https://doi.org/10.1007/s11207-011-9842-2

Sellers, F. B., & Hanser, F. A. (1996). Design and calibration of the GOES-8 particle sensors: The EPS and HEPAD. In E. R. Washwell (Ed.), *Goes-8 and beyond* (Vol. 2812, pp. 353–364). SPIE. https://doi.org/10.1117/12.254083

Shea, M. A., & Smart, D. F. (2012). Space weather and the ground-level solar proton events of the 23rd solar cycle. *Space Science Reviews*, *171*(1), 161–188. https://doi.org/10.1007/s11214-012-9923

Smith, C. W., Heureux, J. L., Ness, N. F., Acuña, M. H., Burlaga, L. F., & Scheifele, J. (1998). The ACE magnetic fields experiment. C. W. SMITH, J. L'HEUREUX and N. F. NESS. *Space Science Reviews*, 613–632. https://doi.org/10.1007/978-94-011-4762-0_21

Stone, E. C., Cohen, C. M. S., Cook, W. R., Cummings, A. C., Gauld, B., Kecman, B., et al. (1998). The solar isotope spectrometer for the advanced composition explorer. *Space Science Reviews*, *86*(1–4), 357–408. https://doi.org/10.1023/A:1005027929871

Tylka, A. J., Cohen, C. M. S., Dietrich, W. F., Lee, M. A., Maclennan, C. G., Mewaldt, R. A., et al. (2005). Shock geometry, seed populations, and the origin of variable elemental composition at high energies in large gradual solar particle events. *The Astrophysical Journal*, *625*(1), 474–495. https://doi.org/10.1086/429384

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al., SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Xiong, B., Wan, W., Zhao, B., Yu, Y., Wei, Y., Ren, Z., & Liu, J. (2014). Response of the American equatorial and low-latitude ionosphere to the x1.5 solar flare on 13 September 2005. *Journal of Geophysical Research: Space Physics*, *119*(12), 10336–10347. https://doi.org/10.1002/2014JA020536

Yashiro, S., & Gopalswamy, N. (2008). Statistical relationship between solar flares and coronal mass ejections. *Proceedings of the International Astronomical Union*, *4*(S257), 233–243. https://doi.org/10.1017/S1743921309029342

Yasyukevich, Y., Astafyeva, E., Padokhin, A., Ivanova, V., Syrovatskii, S., & Podlesnyi, A. (2018). The 6 September 2017 x-class solar flares and their impacts on the ionosphere, GNSS, and HF radio wave propagation. *Space Weather*, *16*(8), 1013–1027. https://doi.org/10.1029/2018SW001932

Zank, G. P., Kryukov, I. A., Pogorelov, N. V., Shaikh, D., Maksimovic, M., Issautier, K., et al. (2010). The interaction of turbulence with shock waves. *AIP Conference Proceedings*, *1216*(April 2010), 563–567. https://doi.org/10.1063/1.3395927