








# Statistical Study of the Correlation between Solar Energetic Particles and Properties of Active Regions

RUSSELL D. MARROQUIN <sup>1,2</sup> VIACHESLAV SADYKOV <sup>2</sup> ALEXANDER KOSOVICHEV <sup>3,4</sup> IRINA N. KITIASHVILI <sup>4</sup>  
VINCENT ORIA,<sup>5</sup> GELU M. NITA <sup>3</sup> EGOR ILLARIONOV <sup>6,7</sup> PATRICK M. O'KEEFE,<sup>5</sup> FRILA FRANCIS,<sup>5</sup>  
CHUN-JIE CHONG,<sup>5</sup> PAUL KOSOVICH,<sup>3</sup> AND AATIYA ALI <sup>2</sup>

<sup>1</sup>*Department of Physics, University of California San Diego, La Jolla, CA 92093, USA*

<sup>2</sup>*Physics & Astronomy Department, Georgia State University, Atlanta, GA 30303, USA*

<sup>3</sup>*Physics Department, New Jersey Institute of Technology, Newark, NJ 07102, USA*

<sup>4</sup>*NASA Ames Research Center, Moffett Field, CA 94035, USA*

<sup>5</sup>*Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102, USA*

<sup>6</sup>*Department of Mechanics and Mathematics, Moscow State University, Moscow, 119991, Russia*

<sup>7</sup>*Moscow Center of Fundamental and Applied Mathematics, Moscow, 119234, Russia*

## ABSTRACT

The flux of energetic particles originating from the Sun fluctuates during the solar cycles. It depends on the number and properties of Active Regions (ARs) present in a single day and associated solar activities, such as solar flares and coronal mass ejections (CMEs). Observational records of the Space Weather Prediction Center (SWPC NOAA) enable the creation of time-indexed databases containing information about ARs and particle flux enhancements, most widely known as Solar Energetic Particle events (SEPs). In this work, we utilize the data available for Solar Cycles 21-24, and the initial phase of Cycle 25 to perform a statistical analysis of the correlation between SEPs and properties of ARs inferred from the McIntosh and Hale classifications. We find that the complexity of the magnetic field, longitudinal lo-

cation, area, and penumbra type of the largest sunspot of ARs are most correlated with the production of SEPs. It is found that most SEPs ( $\approx 60\%$ , or 108 out of 181 considered events) were generated from an AR classified with the 'k' McIntosh subclass as the second component, and some of these ARs are more likely to produce SEPs if they fall in a Hale class with  $\delta$  component. It is confirmed that ARs located in the western hemisphere produced the most SEPs recorded on the Earth's orbit. The resulting database containing information about SEP events and ARs is publicly available and can be used for the development of Machine Learning (ML) models to predict the occurrence of SEPs.

*Keywords:* Sun: activity – sunspots – Sun: particle emission – solar–terrestrial relations

## 1. INTRODUCTION

Solar Energetic Particle (SEP) events are widely known as particle flux enhancements measured by near-Earth satellites. They are among the most dangerous transient phenomena of solar activity because of their negative impacts including health risks for astronauts and airline crews and passengers, damages to satellites and aircraft, radio wave disturbances, power grid disruptions, etc. (Kataoka et al. 2018; Martens 2018). Prediction of SEP events before their occurrence could help diminish their impacts by taking measures ahead of time.

The use of Machine Learning (ML) to predict SEP events has grown significantly in the past years (Sadykov et al. 2021; Torres et al. 2022; Kasapis et al. 2022), in part due to the increase in the readiness of data related to these events. As an example, in the recent review of 36 SEP prediction models by Whitman et al. (2022), 10 models involved ML. Because the performance of ML may be proportional to the quality and availability of data (Goodfellow et al. 2016), we place significant importance on the development of databases with information about SEP events and Active Regions (ARs) expanded in time. ARs are sunspots or sunspot groups that represent regions of strong magnetic fields on the solar surface. They are the primary sources of solar flares and coronal mass ejections (CMEs), and consequently, the SEPs (Reames 2021; Toriumi et al. 2017). Therefore, it is important to study the link between SEP productivity and the basic properties of ARs, which include their location on the solar disk, their areas, various

magnetic field characteristics, and other measures of AR structure and complexity. Probably the longest available observational measures of the AR structure and complexity are the McIntosh (McIntosh 1990) and Hale classifications (Hale et al. 1919). In these classifications, higher classes are assigned to larger and more complex ARs, which are typically more productive in terms of flares, CMEs, and SEPs.

The correlations between SEPs and other solar transient phenomena (flares and CMEs) and McIntosh and Hale classifications were previously studied in several works. For example, Bronarska & Michalek (2017) studied 84 SEP events recorded during the Solar and Heliospheric Observatory (SOHO) spacecraft era (1996-2014) with ARs characterized by the McIntosh classification and found that the most energetic SEPs are ejected only from the associated ARs having a large and asymmetric penumbra. Toriumi et al. (2017) analyzed the Hale classification of ARs that govern large solar flares and eruptions observed between May 2010 and April 2016. McCloskey et al. (2016) studied the flaring rates and the evolution of ARs in terms of the McIntosh classes using data for Solar Cycle 22 (SC22). Their results supported the hypothesis that injection of magnetic energy by flux emergence, which results in an increase in the AR class in the McIntosh system, leads to a higher frequency and magnitude of flare events and, thus, higher SEP event production. The operational flare and SEP daily probabilistic forecasting efforts at the Space Weather Prediction Center (SWPC) at the National Oceanic and Atmospheric Administration (NOAA) also rely on the AR classes (Bain et al. 2021)

The studies mentioned above confirmed the importance of the AR classification and the need for long cross-cycle studies of AR properties and related transient activity. The primary goals of this paper are (1) to develop a homogeneous 40-year-long period data set of Hale and McIntosh classes of active regions and associated SEP events spanning from December 1981 to December 2021, which covers the declining phase of SC21, SC22-24, and the rising phase of SC25, and (2) to perform a direct statistical study of the correlations between SEPs and the Hale and McIntosh classes of ARs using these data. The catalogs utilized in this study are the Solar Region Summary (SRS) records available from the SWPC NOAA<sup>1</sup> starting from 1996, the United States Air Force (USAF) records of AR classes<sup>2</sup> available from 1981 until 2017, and the

<sup>1</sup> <ftp://ftp.swpc.noaa.gov/pub/warehouse/>

<sup>2</sup> [https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/sunspot-regions/usaf\\_mwl/](https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/sunspot-regions/usaf_mwl/)

catalog of the solar energetic particle events affecting the Earth maintained by NOAA<sup>3</sup> with the information about proton events available since 1976. The paper is structured as follows. Section 2 describes the McIntosh and Hale classifications of ARs. Section 3 describes the catalogs used in this study and related data preparation steps. Section 4 highlights the results of the statistical analysis of the association of AR classes and properties with SEPs, followed by the summary of our findings in Section 5. The generated homogeneous data set of AR properties spanning from 1981 until 2021 is publicly available at the Solar Energetic Particle Prediction Portal (SEP<sup>3</sup>) webpage<sup>4</sup>.

## 2. DESCRIPTION OF ACTIVE REGION CLASSIFICATION

### 2.1. *Hale Classification of Active Regions*

The Mount Wilson (or Hale) classification system for sunspot groups put forward by Hale et al. (1919) has been used for nearly a century. This type of magnetic classification provides a simple way to describe the configuration of the magnetic flux and sunspots in an AR (Jaeggli & Norton 2016). According to this classification,  $\alpha$  is a unipolar sunspot group configuration,  $\beta$  is a distinct bipolar configuration with opposite magnetic polarities,  $\gamma$  is a complex configuration with irregular distribution of polarities, and  $\delta$  is a configuration with a sunspot umbra that contains opposite magnetic polarities separated by less than  $2^\circ$  within one penumbra. When appropriate, the classification can include combinations of the primary classes. For example,  $\beta\gamma$  is a complex bipolar configuration with more than one continuous line connecting the opposite polarities. Besides that, if the sunspot group configurations contain one or more  $\delta$  spots, the  $\delta$  class is added to the  $\beta$ - $\gamma$  classification (He et al. 2021). Here we summarize the classes and subclasses presented in the USAF data set:

1. ALPHA ( $\alpha$ ). A single sunspot, or a unipolar sunspot group, around which the distribution of plage is fairly symmetrical. Magnetic field measurements show that the unipolar groups are often accompanied by an area of opposite polarity in which sunspots are not visible.

<sup>3</sup> <https://umbra.nascom.nasa.gov/SEP/>

<sup>4</sup> <https://sun.njit.edu/SEP3/datasets.html>

- 1.1. ALPHA p ( $\alpha$ ). The magnetic field polarity in and around the spot(s) corresponds to the polarity of the leader spots in the same hemisphere for the current solar cycle. The spot(s) and adjacent plage are followed by an elongated area of plage or faculae of the opposite polarity (used in the USAF classification only).
- 1.2. ALPHA f ( $\alpha$ ). The magnetic field polarity in and around the spot(s) corresponds to the polarity for the trailer spots in the same hemisphere for that cycle. The spot(s) and adjacent plage are preceded by an elongated area of plage or faculae of the opposite polarity (USAF only).
2. BETA ( $\beta$ ). A bipolar group in which magnetic field strengths and spot areas indicate a balance between the leader and trailer spots. The polarities show a clear separation.
  - 2.1. BETA p ( $\beta$ ). A bipolar group, in which the magnetic field strengths and spot areas indicate that the leader spot is dominant (USAF only).
  - 2.2. BETA f ( $\beta$ ). A bipolar group, in which the magnetic field strengths and spot areas indicate that the trailer spot is dominant (USAF only).
3. BETA-GAMMA ( $\beta\gamma$ ). A spot group that has  $\beta$  characteristics, but is lacking a well-defined dividing line between regions of opposite polarity. This class includes cases in which spots of the opposite or “wrong” polarity accompany the leader or trailer regions.
4. GAMMA ( $\gamma$ ). A spot group in which the polarities are completely intermixed.
5. BETA-DELTA ( $\beta\delta$ ). A spot group, which has  $\beta$  characteristics, but has umbrae of opposite polarity inside the same penumbra.
6. BETA-GAMMA-DELTA ( $\beta\gamma\delta$ ). A spot group, which has  $\beta\gamma$  characteristics, but has umbrae of opposite polarity inside the same penumbra.
7. GAMMA-DELTA ( $\gamma\delta$ ). A spot group, which has  $\gamma$  characteristics, but has umbrae of opposite polarity inside the same penumbra.
8. DELTA ( $\delta$ ). A sunspot group with umbra having opposite polarities within a penumbra **that** spans less than two heliographic degrees.

## 2.2. McIntosh Classification of Active Regions

The McIntosh classification scheme was originally developed by Cortie (1901) and later modified and expanded to include a wider range of parameters by McIntosh (1990). It describes the white-light structure of sunspot groups and is composed of 60 allowed classification combinations derived from 17 different parameters (McCloskey et al. 2016). The general form of the McIntosh classification is  $Zpc$ , where  $Z$  is the modified Zurich class,  $p$  is the type of largest spot, and  $c$  is the degree of compactness in the interior of the group. A more detailed description of the classification scheme based on the USAF database documentation is provided below.

### 2.2.1. Modified Zurich Class – $Z$

The modified Zurich classes are defined on the basis of whether penumbra is present, how the penumbra is distributed, and by the extent of the group (McIntosh 1990). In contrast to the original Zurich definitions, a judgment of complexity is not required. There are seven classes in this component of the system described below.

- A* Unipolar group with no penumbra with the total extent (normally) of less than 3 heliographic degrees.
- B* Bipolar group of spots with no penumbra; the length is (normally) 3 heliographic degrees or greater.
- C* Bipolar group of spots when only spots of one polarity have penumbra, usually the spots at one end of an elongated group.
- D* Bipolar group when spots of both polarities have penumbra. The group length does not exceed 10 heliographic degrees.
- E* Bipolar group when spots of both polarities have penumbra. The group length is greater than 10 but less than or equal to 15 heliographic degrees.
- F* Bipolar group when spots of both polarities have penumbra. The group length exceeds 15 heliographic degrees.

*H* Unipolar group of spots with penumbra. The principal spot is usually the leader spot remaining from an old bipolar group.

### 2.2.2. *Penumbra of Largest Spot – p*

The type of largest spot in a sunspot group can be described by a combination of type, size, and symmetry of penumbra and umbrae within a given penumbra (McIntosh 1990). There are six classes in this component described below.

*x* No penumbra.

*r* Rudimentary or incomplete irregular penumbra. It is brighter than a mature penumbra and has a mottled or granular (not filamentary) fine structure.

*s* Small symmetric penumbra. Mature, dark, circular, or elliptical penumbra with a filamentary fine structure; the diameter across the penumbra is 2.5 heliographic degrees or less. This class includes penumbrae that appear elliptical due to the effect of geometric foreshortening. Symmetric penumbra usually contains either a single umbra or a compact cluster of umbrae near the sunspot center.

*a* Small asymmetric penumbra. Mature, dark, irregular (clearly not circular or elliptical) penumbra with filamentary fine structure; the diameter across the penumbra is 2.5 heliographic degrees or less. The asymmetry is “real”, not just due to foreshortening effects. An asymmetric penumbra usually contains two or more umbrae scattered within it.

*h* Large symmetric penumbra. It has the same characteristics as a small symmetric(s) penumbra; the diameter across the penumbra is greater than 2.5 heliographic degrees (normally corresponding to an area greater than about 250 millionths of the solar hemisphere).

*k* Large asymmetric penumbra. It has the same characteristic as a small asymmetric (a) penumbra; the diameter across the penumbra is greater than 2.5 heliographic degrees (normally corresponding to an area greater than about 250 millionths of the solar hemisphere).

### 2.2.3. *Sunspot Distribution – c*

This component of the three-letter classification indicates the density of an internal spot population in a sunspot group. A ranking of spot distribution in the interior of a sunspot group gives additional information about the area of the group and the potential presence of strong spots near the line of polarity inversion lying between the principal leader and follower spots (McIntosh 1990).

- x* Undefined for a single sunspot or unipolar spot group.
- o* Open. Few, if any, spots between the leader and trailer spots. Any interior spots are very small umbral spots or pores.
- i* Intermediate. Many spots lie between the leading and trailing portions of the group, but none of them possesses mature penumbra.
- c* Compact. The area between the leading and trailing ends of the spot group is populated with many strong spots, with at least one interior spot possessing mature penumbra. An extreme case has the entire spot group enveloped in one continuous penumbral area.

A summary of the McIntosh classification is given in Table 1. The total number of possible classes is 60. It is important to note that not every combination of the components mentioned above is permitted.

Class	Penumbra: Largest Spot	Distribution	Number of Configurations
A	x	x	1
B	x	o,i	2
C	r,s,a,h,k	o,i	10
D,E,F	r	o,i	6
D,E,F	s,a,h,k	o,i,c	36
H	r,s,a,h,k	x	5
Total allowed types	–	–	60

**Table 1.** McIntosh Class Configurations of Sunspot Groups.



### 3. DATA PREPARATION

#### 3.1. *Homogeneous Data Set of Solar Active Regions*

The USAF dataset of active region records contains information about AR classes from December 1981 to December 2017. The currently-maintained SWPC NOAA SRS records range from January 1996 to the current time. To maximize the availability of data for this statistical analysis, we combine the USAF and SWPC NOAA AR catalogs into a continuous AR database covering the 40-year period, from December 1981 to December 2021. However, the catalogs are not entirely consistent in the way the AR classes are reported. For example, SWPC NOAA SRS records contain information about ARs once per day at 00:00 UT, while the USAF dataset can contain several records of the same AR throughout the day from different observing sites at different times. Moreover, many solar energetic particle events are originating from the regions close to the western limb where the information about AR is either not available or ambiguous. Therefore, we have performed certain steps toward the homogenization of the data sets, namely bringing the records from the USAF data set into a form compatible with the current SWPC NOAA SRS reporting.

The USAF catalog files were held separately for each year and in a text file format. Relevant information to perform this statistical analysis was acquired line by line and character by character, following the documentation provided for this data catalog. In order to keep consistency with the SWPC NOAA records and the accuracy of its contents, these records were edited in accordance with the documentation provided by the USAF. The years provided in a YY format were changed to YYYY format so that the dates followed the consistent format *YYYY-MM-DD hour:min:sec*. According to the documentation, the location of an observed AR is given respectively by six characters in the following order: E or W for East or West, two integers for longitude given as central meridian distance, N or S for North or South, and two integers for heliographic latitude. The negative sign was added for the South and East longitudes.

The records from the USAF came from four different observational stations, such as the Mount Wilson Observatory and the Boulder Observatory among others. Following the documentation, the area of each AR is given in millionths of a solar hemisphere. It was noticed that some entries of the area were equal to zero, while additional entries for the same AR from other observatories had nonzero areas. We decided that if the area of an AR was equal to zero while being recorded as nonzero by any other observatory,

such an AR would take its respective and most recent nonzero area value, independently from which observatory the information was obtained. This method was carried out through an algorithm by first sorting the respective dataset by AR number, date, and time of observation, which resulted in groups of ARs that ascended based on the parameters just mentioned. As our algorithm iterated through each entry if an entry with a value of area equal to zero was reached, it inspected the previous or next most recent records of the respective AR. If the algorithm found a nonzero value for the same AR, it replaced the zero value and moved on to the next entry.

According to the documentation, the USAF dataset followed the Hale classification scheme described in Section 2.1. A variety of inconsistencies were found throughout the catalogs for each year, and such inconsistencies were updated accordingly. Hale classifications were given as single letters with up to four character spaces, depending on the sunspot group configuration, and were changed accordingly. For instance,  $A$  was replaced by  $\alpha$ ,  $B$  was replaced by  $\beta$ ,  $BGD$  was replaced by  $\beta\gamma\delta$ , and so on for every entry. Subclasses of  $\beta$  and  $\alpha$  (items 1.1, 1.2, 2.1, and 2.2 in Section 2.1) were designated as  $\beta$  and  $\alpha$ , respectively. Following the sorting method based on the AR number, date, and time, an algorithm iterated through every entry. When a Hale class was considered ambiguous, meaning that its magnetic type was unclear, we inspected the previous or next most recent records of the respective AR. If the algorithm found an appropriate magnetic type for the same AR, it replaced the Hale class in question and moved on.

The McIntosh classification scheme described in Section 2.2 is inferred from McIntosh (1990) and the USAF documentation. The three-component McIntosh classes given by three characters were extracted one by one from the USAF datasets. Inconsistencies in these records were divided into two categories: “ambiguous” and “unambiguous”. The McIntosh classes were designated ambiguous when one of the three components was missing, and this missing component had more than one allowed class outlined in Table 1. For instance, the McIntosh classification  $'FI'$  was labeled as ambiguous because the penumbra type of the largest spot (second component) was not specified, and based on table 1, several different classes in the second component are allowed to be paired with  $'F'$  class (first component) and  $'I'$  distribution of sunspots (third component). In addition, several McIntosh configurations contained one or more components paired with one or more disallowed components. To illustrate, the McIntosh configurations starting

with the 'H' class as the first component and any penumbra type of the largest spot (listed in Table 1) were considered as "ambiguous" because five different classes are allowed to take its place as the second component. Nevertheless, any ARs classified with 'H' for the first component whose third component is inconsistent can be considered "unambiguous" because only the 'X' class is allowed in place of such inconsistency. As a result, the McIntosh Classification 'HXO' was labeled as "ambiguous" because of the following reasons: 1) The 'H' class is not allowed to be paired with the second component 'X' or the third component 'O' as defined in Table 1. 2) Although the third component can be unambiguously edited to be 'X', there exist five different allowed classes for the second component, leaving no unambiguous choice for such a component.

In contrast to the "ambiguous" McIntosh classifications, an "unambiguous" classification refers to a McIntosh configuration with specifically one or more missing components, which can be unambiguously added to the respective record. A caveat of the last sentences is that the word "specifically" is important because, for example, the McIntosh classification 'AXO' could easily be considered as "unambiguous" since the third component can be unambiguously replaced with 'X' to form 'AXX'. Nevertheless, according to Table 1, the first component 'A' can be edited to change the configuration from 'AXO' to 'BXO', which is also an allowed configuration. This peculiarity makes 'AXO' an "ambiguous" McIntosh configuration. An example of an unambiguous McIntosh configuration is 'AX' with a missing classification of the distribution of the AR in consideration. This was considered unambiguous because the distribution component is missing and only the 'X' class is allowed as the third component for this McIntosh configuration (as shown in Table 1), while the first two components followed an allowed configuration.

After defining the different inconsistencies found in the records of McIntosh classifications, and following the sorting method based on the AR number, date, and time, an algorithm iterated through each entry with information about ARs for a given year. Because the datasets from each catalog were previously sorted according to the AR number, date, and time of observation, when the algorithm found an ambiguous entry, it inspected the most recent previous or next record of the same AR. If an acceptable McIntosh classification was found, referring to a neither "ambiguous" nor "unambiguous" McIntosh class, such a

class replaced the McIntosh class in question, irrespective of which observatory the record was acquired. The unambiguous McIntosh configurations were found individually and corrected accordingly.

In order to keep the data formatting consistent throughout the merged AR database, each AR record in the USAF catalogs had to be approximated to midnight, similar to the SWPC NOAA SRS records. Following the sorting method previously used for editing specific records of ARs, our algorithm iterated through each entry in the catalog for a given year and approximated its day and time to the next midnight based on four conditions:

1. If the next entry has the same day and same AR number, store its index and continue to the next entry.
2. If the next entry has the same day and a different AR number, approximate the current entry to midnight.
3. If the next entry has the same AR as the next day, approximate the current entry to midnight.
4. If the last entry is reached, approximate this entry to midnight.

We approximate and keep only the last record of an AR during a day, independently from which observatory the record was noted. *Condition 1* ensures that if the next AR record has the same date, meaning that such a record is not the last observation of the same AR during a given day, the corresponding index will be saved. Later, saved indices were dropped, keeping only the last and the approximated record of each respective AR during each day. When the first condition was not fulfilled, *Condition 2* examined if the next AR record is from a different AR. If true, this meant that the current AR record was its last observation of the day, since AR records were sorted in ascending order of AR number, date, and time, and it was approximated to midnight. If previous conditions were not met, it meant that the next record had the same AR number with a different date. *Condition 3* determined whether the next AR record corresponded to another day. If true, it meant that the current record was the last observation of the respective AR during the current day, and such an AR record was approximated to midnight. Because the last entry of the sorted catalogs for each corresponds to the last observation of any AR during that day, *Condition 4* approximates such a record to midnight.

Following the time approximation of the respective ARs, it was natural to also approximate their longitudinal location according to the Carrington rotation rate using the following formula:

$$\text{New Longitude} = \text{Current Longitude} + \left( \frac{\text{difference in hours until midnight}}{24 \text{ hrs}} \right) * 14.2^\circ \quad (1)$$

By making the modifications described above, we were able to construct the continuous AR dataset by combining the data catalogs from the SWPC NOAA and USAF for December 1981 - February 2021. This dataset is presented in Figure 1(a). We have also tested the homogeneity of the data provided by the two sources we utilize after applying the processing steps above to AFRL records. This was done through quantitative comparisons of the number of ARs and their respective features during several overlapping years, from January 1, 1997, to December 31, 2004, during the SC23 maximum. Figure 2 shows the histograms used for our homogeneity test. The histogram on the left was generated using data solely from the SWPC NOAA database, while the histogram on the right was produced using data from the USAF database. Although the total number of ARs in each dataset is relatively close, the height of several bins corresponding to a specific period of time seems to differ, meaning that the number of ARs recorded by the two solar centers for every single time is not exactly the same. In addition, the test showed that the portion of ARs with  $\beta\gamma\delta$  (brown),  $\beta\gamma$  ARs (red), and  $\beta$  (orange) magnetic field types also slightly differ in these databases. Thus, we determined that the datasets obtained from the two different solar centers while generally consistent are not entirely homogeneous, and this has to be taken into account while utilizing the developed dataset for research purposes.

### 3.2. Linking Records of Solar Energetic Particle (SEPs) and Active Regions

The SWPC NOAA list of SEP events affecting the Earth provides records of the SEP events that, according to the observations of the Geostationary Operational Environmental Satellite (GOES), reached the threshold of 10 particle flux units (pfu) for  $\geq 10$  MeV protons. The current records of SEP events in our possession span six decades, starting from April 1976 until September 2017. For the complete list of the SEP events please refer to the original source, <https://umbra.nascom.nasa.gov/SEP/>. The list contains information about the SEP event start and peak time, the peak flux of  $\geq 10$  MeV protons, and the corresponding

information about the preceding Coronal Mass Ejection (CME) and soft X-ray flare record. The majority of the records also contain information about the location of the host active regions on the solar surface and their NOAA number, allowing us to link this list with the homogeneous AR dataset constructed above. We also note here that, although the SEP records provide their start and peak date and time, we utilize the time of the preceding flare (more precisely, its peak in 1-8 Å soft X-ray emission) as a reference time for merging the datasets, because the SEP arrival time may vary from minutes to several hours.

Inconsistencies were also found in the SEP records and were updated to undertake this statistical analysis. The total number of SEP records in the analyzed list is currently 267. From that number, 36 records did not have the flare maximum date and time and were removed, leaving 231 SEPs. Among these, 11 records did not have AR numbers and were also removed. We use the remaining 220 SEP events to study statistical relationships between ARs and SEPs.

Not all SEPs originated from ARs observed on the visible solar disk, or within the  $[-90^\circ, 90^\circ]$  longitude range. Figure 3(a) shows that several SEP events were identified as originating from the ARs behind the western limb with a longitude  $> 90^\circ$ . The western hemisphere is more directly magnetically connected to the Earth (Parker 1958) and the statistical studies of SEP origins demonstrated the asymmetry towards the western limb (Cliver et al. 2020). Therefore, it is important to include information about the ARs located close to the western limb and to map some of the SEP events to these ARs. To accommodate for such SEPs, we decided to extrapolate the longitude of all ARs (and their corresponding dates) to cover the entire  $360^\circ$  circumference of the Sun by applying the Carrington rotation rate. In order to avoid any duplicates, the AR longitudes were extrapolated only after their last records near the west limb. The extrapolation was performed assuming the Carrington rotation rate following Eqn. 1. Figure 1(b) illustrates the continuous AR database after performing the extrapolation. It can be interpreted as a stacked histogram with the total number of extrapolated Active Regions per year. The legend shows the Hale classifications and the contribution of Active Regions (ARs) of a respective Hale class to the total number of ARs in each bin, corresponding to a range of dates. It can be observed from each bin that the number of ARs increased approximately by a factor of three compared to the histogram of ARs in Figure 1(a).

#### 4. RESULTS AND DISCUSSION

After acquiring the extrapolated AR database and the SEP database, we merged the records into a single dataframe containing a one-to-one correspondence between ARs and SEPs generated by ARs. The merging process was performed by considering the corresponding AR numbers and the date of the consideration of an AR with the date of the peak flare time record from a SEP event (i.e., the class of the AR recorded in the preceding midnight with respect to the flare peak time was linked to the flare and, correspondingly, to the SEP records). From the 220 SEP events selected, as described in Section 3.2, we were able to match 181 SEPs to their AR sources in our extrapolated AR database.

#### 4.1. *Properties of Active Regions*

Figure 3(a) shows the total number of SEP events vs. longitude. The legend shows the contribution of ARs in a respective Hale classification to the number of SEP events within each range of longitude. It can be observed that most SEPs are produced in the Western Hemisphere of the Sun, within the range  $[0, 90]$ . This characteristic arises because the Earth is magnetically connected to the solar longitudes of  $\approx 75^\circ$ , meaning that SEPs are more likely to reach Earth if it originates from an AR closer to that range of longitudes. Although this relationship was known and highlighted in the previous studies (e.g., Cliver et al. 2020), there are some additional interesting dependences related to the magnetic classes of SEP-productive ARs that we can mention.

Figure 3(b) displays the rate of SEP event generation (i.e., a daily climatological probability of the SEP to be produced from the AR of a certain class) in four longitude bins:  $(-81^\circ, -35^\circ)$ ,  $(-35^\circ, 12^\circ)$ ,  $(12^\circ, -58^\circ)$ , and  $> 58^\circ$ . The daily climatological probability is calculated as the number of SEPs generated from ARs in a respective Hale class divided by the total number of ARs in our AR database with the same respective Hale class multiplied by 100%, taking into account that each AR has only one record daily. One can notice that the probabilities increase towards the western longitudes for both the simpler configurations of ARs (like  $\beta$ ) and more complex  $\beta\gamma\delta$ . It also is inferred from Figure 3(a) that a greater number of SEPs came from an  $\alpha$  (blue) AR when they were closer to the Earth and the Sun's magnetic connection, while the distribution of SEPs originating from  $\beta\gamma\delta$  (brown) ARs shows that a relatively large number of SEPs were generated far from  $\approx 75^\circ$ . This can be related to the fact that more complex ARs generate stronger flares (in terms of their soft X-ray class) that statically result in faster and wider CMEs. The CME width was

recently indicated to be an important parameter for SEP forecasting (Torres et al. 2022). On the other hand, the recent work by Laitinen et al. (2023) indicated that SEPs can arrive at a wide range of longitudes, even without a wide particle source. In addition, it can be inferred that SEPs can be generated behind the western limb. In our case, six SEPs were found to be generated from such regions.

Figure 3(c) shows a histogram with the number of SEPs vs. the area of their respective ARs. It can be generally inferred that larger ARs, namely ARs with larger areas, are more likely to produce SEP events because the total number of SEPs increases with increasing area. In addition, we can observe that SEP productivity favors larger ARs regardless of their Hale classification. Essentially, the distribution of SEPs increases as the AR area and their corresponding Hale classes grow.

Figure 3(d) shows the number of SEPs events vs. their respective dates. Comparing with the histograms in Figure 1, we can see that the SEP event productivity varies in phase with the 11-year solar activity cycle. It also demonstrates the known pattern of the solar cycle 24 being weaker than preceding cycles 23 and 22 both in terms of the sunspot numbers, numbers of ARs that appeared at the surface, and the number of generated SEPs.

#### 4.2. Hale Classification

Figure 4 shows the correlation between the Hale classification of ARs (described in Section 2.1) and SEP productivity. Panel (a) shows the percentage of the total number of SEPs produced by ARs classified by a respective Hale class. The percentage was calculated by dividing the total number of SEP events found in a corresponding Hale class by the total number of SEPs in our database, which is 181 SEP events in total. Approximately 33% of the total number of SEPs were produced by  $\beta$  ARs, followed by the  $\beta\gamma\delta$  class with approximately 29%. The  $\delta$  ARs appear to have produced the least amount of SEP events with approximately 1% of the total number of SEPs. Interestingly, even the ARs that are typically assumed to be not active with respect to solar transient events (such as  $\alpha$  or  $\beta$ ) generated more than 40% of the SEPs considered in this work if combined. This confirms the importance of the characterization of ARs with other parameters in addition to the Hale class, such as McIntosh classes or quantitative magnetic field properties (Kasapis et al. 2022; Sadykov et al. 2021).



The  $\delta$ -class ARs and ARs exhibiting the  $\delta$  Hale class in their magnetic field configuration are historically regarded as highly SEP productive because of their correlation with solar flares. Figure 4(b) shows the rate of SEP productions (i.e., the daily climatological probability of SEP events in these regions, shown as a percentage on the y-axis) of ARs in a respective Hale class. It can be inferred that the  $\delta$  configurations have the highest chance of the SEP event generation (approximately 7.1%) given the condition that the  $\delta$  region is observed, confirming what has been historically concluded. The  $\beta\delta$  regions show a rate of approximately 3.2%,  $\beta\gamma\delta$  regions have a rate of approximately 3.5%,  $\gamma\delta$  ARs give approximately 5.6%, and  $\delta$  regions show the highest rate of SEP production at approximately 7.1%.

#### 4.3. McIntosh Classification

Figure 5 shows the results of a statistical study referencing the McIntosh classification of ARs, described in Section 2.2, and SEP events. Panel (a) shows the distribution of the total number of SEPs that were produced by ARs with their respective three-component McIntosh classification. Through the legend, we can distinguish the proportion of ARs from a corresponding Hale class that produced SEPs. Thus, it can be inferred that the largest quantity of SEP events originated from ARs having 'k' as the second McIntosh component. Additionally, the majority of these ARs were classified as  $\beta\gamma\delta$ , which allows us to close in on a connection between the Hale class ( $\beta\gamma\delta$ ) and the McIntosh subclass (k). Figure 5(b) shows the distribution of SEP events generated from ARs with a respective AR class given by the first component of the McIntosh classification scheme, while Figure 5(c) indicates the total number of SEPs which originated from ARs with a respective distribution of Sunspots provided by the first McIntosh component. Because the legend is kept consistent throughout the entire figure, it can be inferred from Figure 5(b) that most SEPs were generated from ARs exhibiting classes E, D, F and corresponding Hale classes  $\beta\gamma\delta$ ,  $\beta\delta$ ,  $\beta\gamma$ , and  $\beta$ . On the other hand, Figure 5(c) shows an almost even total number of SEPs originated from ARs with the distribution of sunspots *i*, *c*, and *o* with a less even proportion of such ARs in a respective Hale class. The two most obvious correlations are the  $\beta\gamma\delta$  regions, which appear to favor the 'c' distribution of sunspots, and the  $\beta$  regions favoring the 'o' counterpart.

A more defined association of SEP events and McIntosh classes is shown in Figure 6(a). It shows the total number of SEP events generated by ARs with a given penumbra type of the largest sunspot in the AR,

defined as the second component of a McIntosh class. Based on this figure, we can conclude that ARs with a 'k' subclass as the second component of a McIntosh classification is the most SEP productive. In addition, it can be inferred that ARs with a  $\beta$  and  $\gamma$  or  $\delta$ , and a  $\beta\gamma\delta$  magnetic-field configurations favor a 'k' subclass as the second component of a McIntosh class in terms of SEP production. Figure 6(b) shows the rate of production of SEPs from ARs with a respective penumbra type of their corresponding largest sunspot. The rate was calculated as calculated in Section 4.2 for Hale classes. The number of SEPs generated from ARs in a respective class is divided by the total number of ARs in our AR database with the same respective class multiplied by 100%. Following the results, it can be concluded that ARs with the 'k' subclass as the second McIntosh component has a higher rate of SEP production, and consequently, they are more probable to produce SEPs than ARs with any other class in this component.

#### 4.4. Combined McIntosh and Hale Classifications

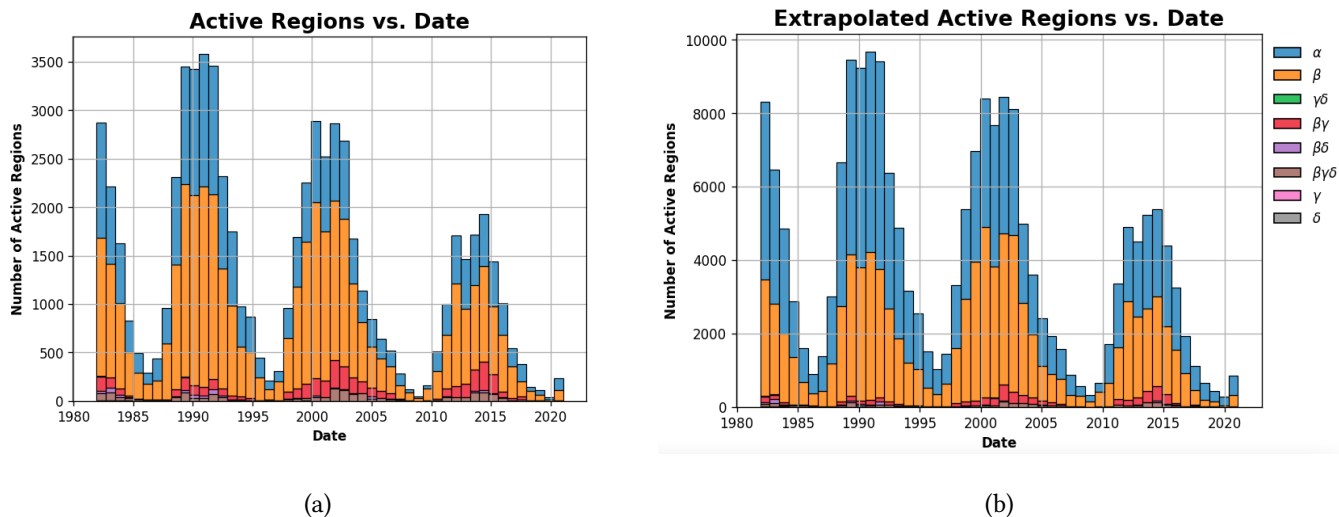
Figure 6 shows that some ARs are more likely to produce SEPs when their second McIntosh component is determined as 'k' and paired with certain Hale class configurations. To further study this relationship, we combined the two classification schemes and ARs were grouped based on two features: the Hale class and the second component of the McIntosh class. The results are shown in Figure 7. We can conclude that the rate of SEP production is the highest for ARs with any given Hale class combined with the 'k' subclass as the second component of a McIntosh classification. In addition, it is important to note that these rates are higher than the rates found by separately considering Hale classes and the considered subclass of the McIntosh classification. For example, the rates of SEP events produced by  $\delta$  ARs and ARs with 'k' second McIntosh components were  $\sim 7.1\%$  and  $\sim 1.8\%$  in that order and as shown in Figures 4(b) and 6(b). Because (k,  $\delta$ ) ARs have an  $\sim 9\%$  rate of SEP production, they are more likely to produce SEPs than the ARs that are classified separately as 'k' or  $\delta$ . The same can be concluded for  $\beta$ ,  $\beta\delta$ ,  $\beta\gamma$ ,  $\gamma\delta$ , and  $\beta\gamma\delta$  ARs if they are also determined to have the 'k' McIntosh subclass. These results indicate the importance of including both Hale and McIntosh classes for advancing the forecasting of SEPs.

## 5. SUMMARY

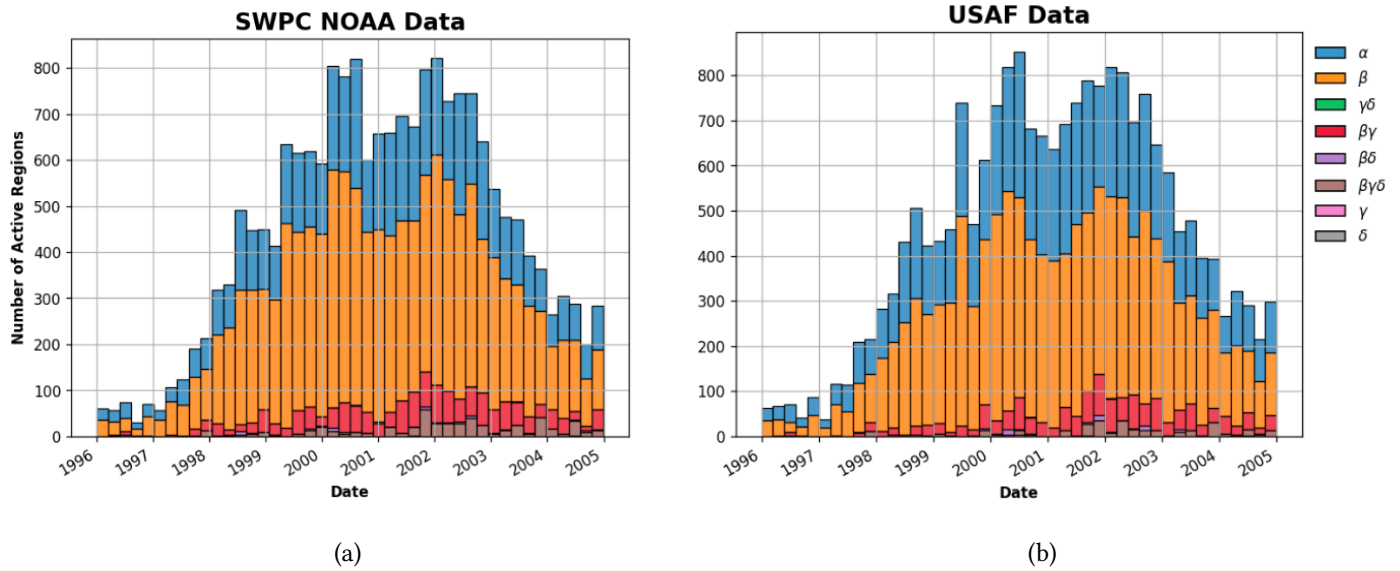
We presented a statistical study of the correlation between SEP events and properties, Hale, and McIntosh classes of ARs. Properties of ARs included longitude and area of ARs, and phases of a solar cycle, while classes came from the Hale and McIntosh classification schemes. We concluded that the longitudes closer to the magnetic connection between the Sun and the Earth are more favorable for less complex ARs, such as  $\alpha$  and  $\beta$  ARs, in terms of SEP production. In addition, it was noted that SEPs may be produced beyond the  $[-90^\circ, 90^\circ]$  longitude range of our Sun's surface. Larger areas of ARs and the amount of solar activity may also be related to SEP production, given our statistical evidence. In addition, we inferred that the total number of SEPs produced by ARs of a certain Hale or McIntosh class is not a sufficient characteristic to determine the hazard a given AR may represent. For instance, although  $\delta$  ARs generated only one SEP event throughout the entire time frame in consideration, they have a higher rate of producing SEP events, given the total number of times such ARs were observed in the same time frame. A more conclusive relationship can be inferred through considerations of the second component of a McIntosh class and the total number of SEPs. It was found that most SEPs were generated from an AR classified with the 'k' McIntosh subclass as the second component ( $\approx 60\%$ , or 108 out of 181 considered events), and some of these ARs are more likely to produce SEPs if they fall in a certain Hale class. For example,  $\delta$  ARs have a  $\sim 7.1\%$  chance to produce SEPs, while ARs with a 'k' as the second McIntosh component has a  $\sim 1.8\%$  rate of SEP production. Nevertheless, an AR has a probability of  $\sim 9\%$  to produce a SEP event if it is determined to have a  $\delta$  Hale classification and whose second McIntosh component is 'k.' The developed homogeneous dataset of AR classes spanning more than three solar cycles (1981-2021) can be utilized for the development of robust SEP events and other transient solar activity forecasting approaches validated over an extensive time period and varying solar activity.

#### ACKNOWLEDGMENTS

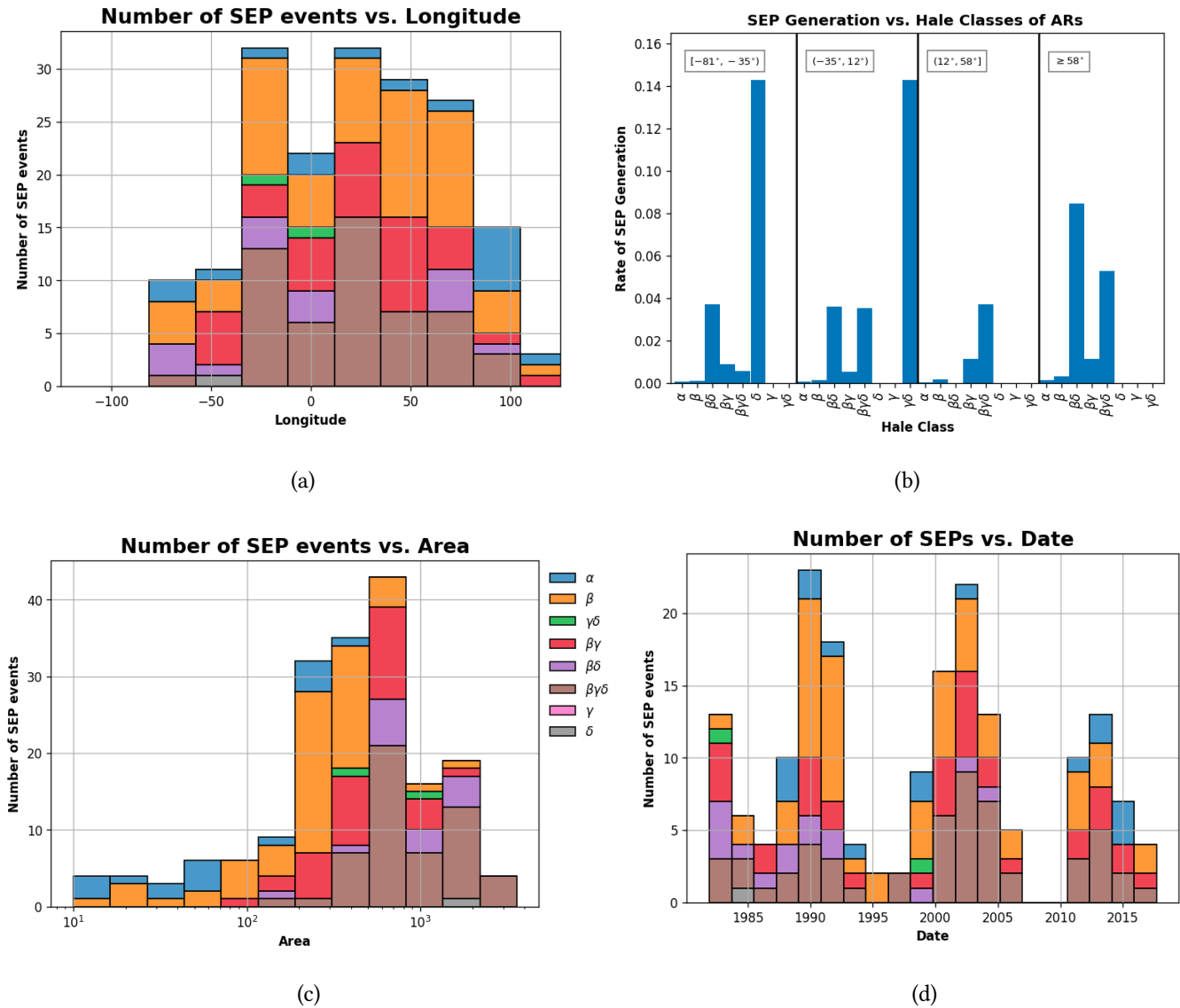
This research was supported by NASA Early Stage Innovation program grant 80NSSC20K0302, NASA LWS grant 80NSSC19K0068, NSF EarthCube grant 1639683, and NSF grant 1835958. VMS acknowledges the NSF FDSS grant 1936361 and NSF grant 1835958. EI acknowledges the RSF grant 20-72-00106.



**Figure 1.** The continuous AR database was achieved by combining Active-Region (AR) catalogs from the Space Weather Prediction Center at the National Oceanic and Atmospheric Administration and US Air Force Space Weather Wing. The annual AR numbers are visualized through a stacked histogram in panel (a). Panel (b) shows a stacked histogram with the total number of extrapolated Active Regions vs. date. The legend shows the Hale classifications and the contribution of ARs in a respective Hale class to the total number of ARs in each bin.

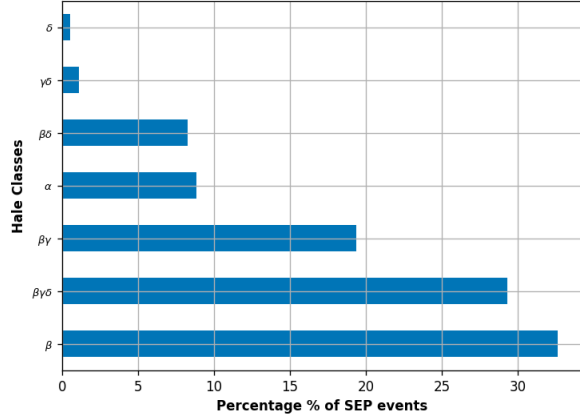


**Figure 2.** Histograms used to perform the homogeneity test. The left panel (a) shows data originating from the Space Weather Prediction Center of the North Oceanic and Atmospheric Administration (SWPC NOAA). Right panel (b) shows data from the US Air Force (USAF). The legend shows the Hale classifications and the contribution of Active Regions (ARs) in a respective Hale class to the total number of ARs in each bin. The appearance of each class is consistent throughout the entire figure.



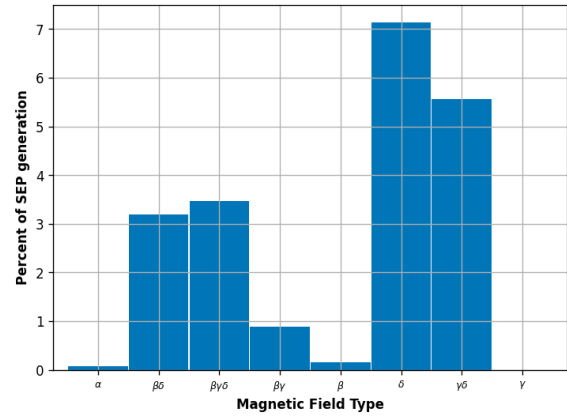
**Figure 3.** Panel (a) shows the number of SEPs vs. the location of ARs along the surface of our Sun. According to this figure, the majority of SEPs were generated in the Western Hemisphere. The legend shows the Hale classifications and illustrates the contribution of ARs in a respective Hale class to the total number of SEPs in each bin. Panel (b) depicts the rate of SEPs produced by Active Regions (ARs) in certain longitude ranges classified by their magnetic field type. Panel (c) shows the number of SEPs vs. the area of their respective ARs. Panel (d) shows the number of SEPs within the specific date ranges in each bin. The legend shows the Hale classifications and illustrates the contribution of ARs in a respective Hale class to the total number of SEPs in each bin.

Hale Classification of ARs vs. Percent of SEP Production



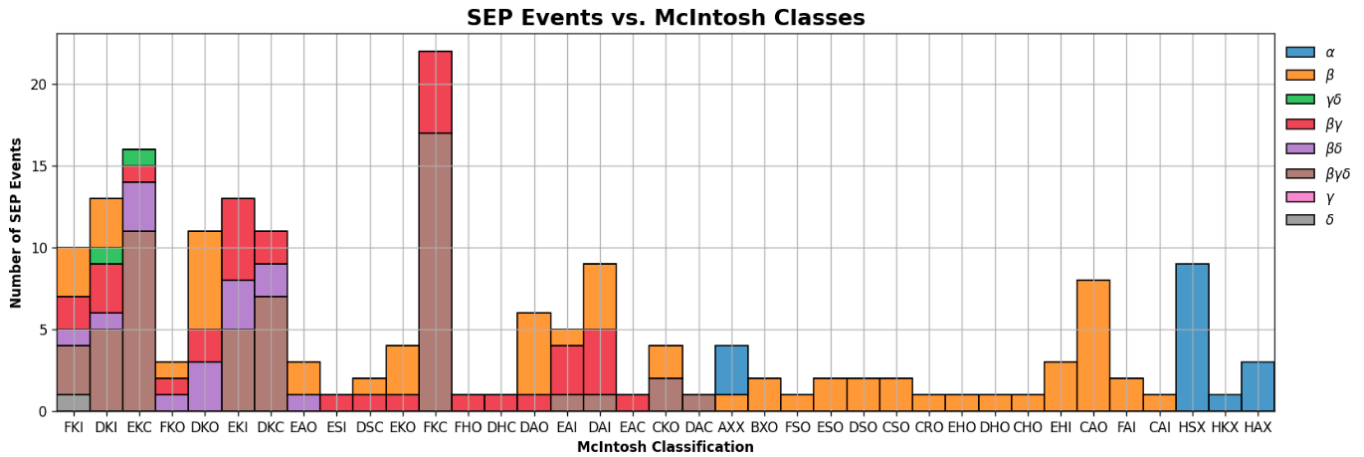
(a)

Rate of SEP events vs. Hale Classes

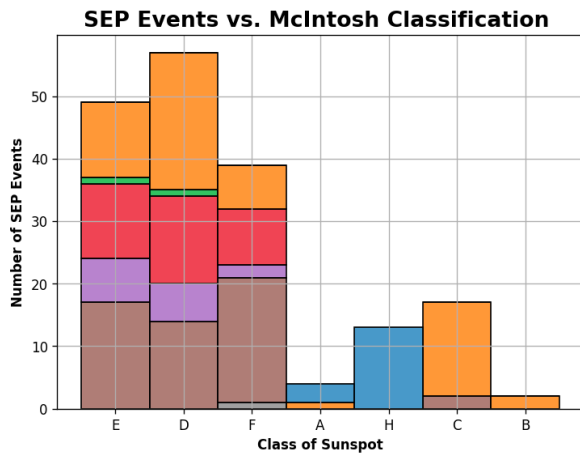


(b)

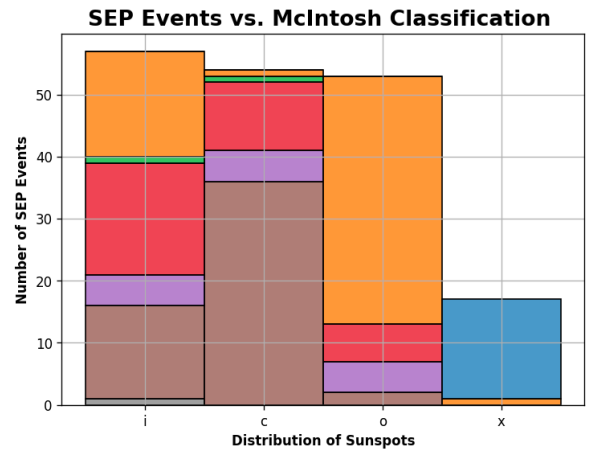
**Figure 4.** Panel (a) shows the fraction of the total number of SEP events (as a percentage) that were produced by ARs with a corresponding Hale class. Panel (b) shows the rate of SEPs produced by ARs with a corresponding Hale class. The rate is calculated by dividing the total number of SEPs produced by an AR from a respective Hale class and its corresponding total number of appearances.



(a)



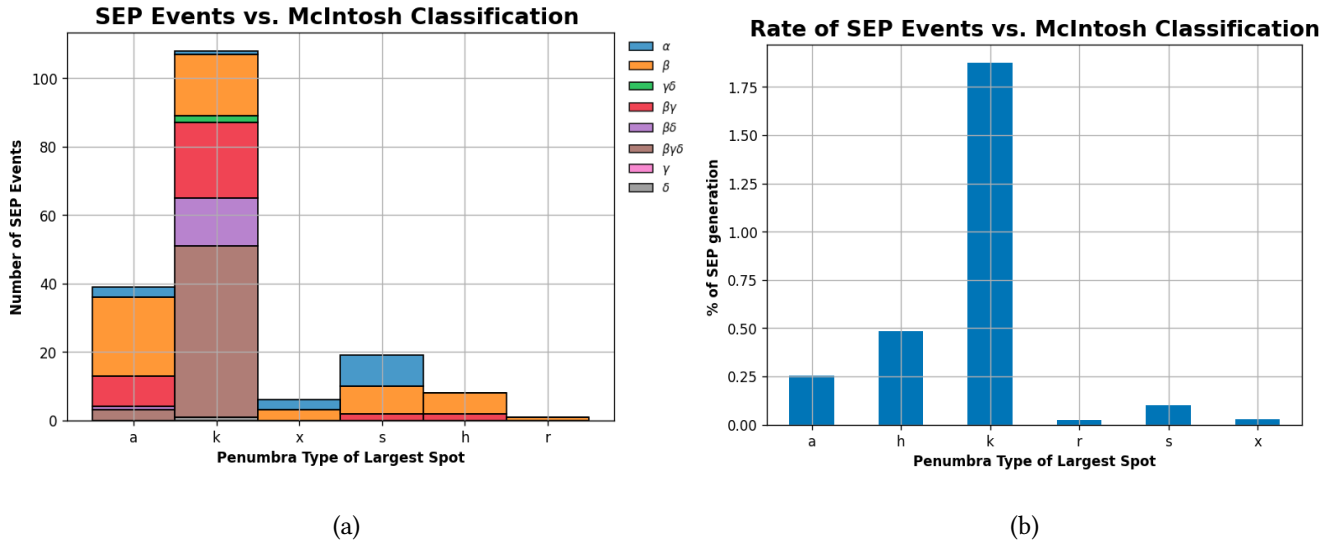
(b)



(c)

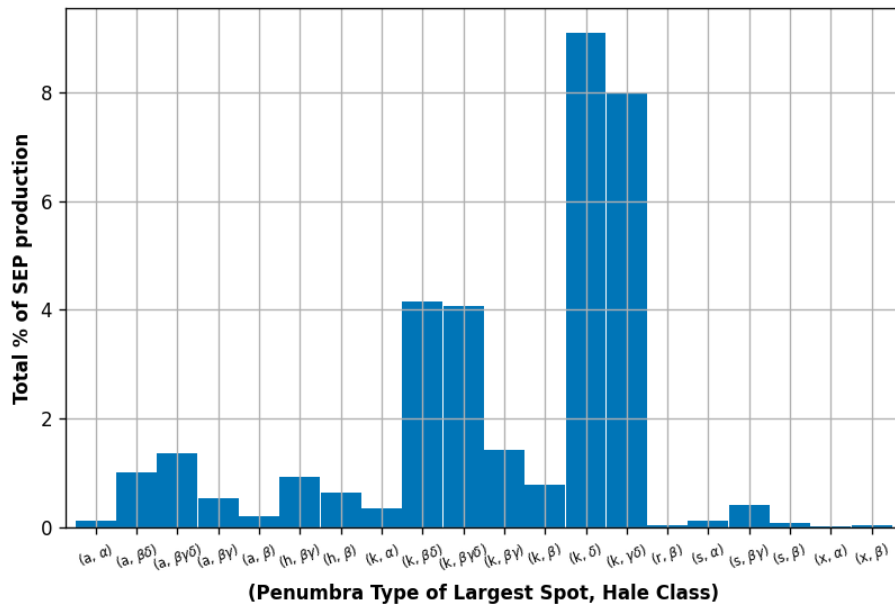
**Figure 5.** Panel (a) shows the distribution of the SEP events over the entire McIntosh Classification scheme. The legend shows the Hale classifications and the contribution of Active Regions (ARs) in a respective Hale class to the total number of SEPs in each bin. The appearance of Hale’s class is consistent throughout the entire figure. After separating each of the McIntosh components, panel (b) shows the number of SEPs produced by ARs classified by the first component of the McIntosh scheme. Panel (c) shows the number of SEPs produced by ARs classified by the third component of the McIntosh component.





**Figure 6.** Panel (a) shows the number of SEP events produced by ARs classified by the second component of the McIntosh classification scheme. The legend shows the Hale classifications and the contribution of Active Regions (ARs) in a respective Hale class to the total number of SEPs in each bin. Panel (b) shows the rate of SEP production of ARs classified by considering the McIntosh component used in panel (a). To find calculate the rate, the total number of SEPs generated from ARs with a respective second McIntosh component is divided by the total number of appearances of ARs with the same corresponding classification.

### Rate of SEP events vs. McIntosh and Hale Classes



**Figure 7.** Rate of SEP events (in percentage) vs. a combination of the second component of the McIntosh classification scheme and the Hale classification scheme. When combined, the rates of SEP production are found to increase for either or both classes in consideration when considered separately. Two major groups of ARs are found to be the most SEP productive:  $(k, \delta)$  and  $(k, \gamma\delta)$  ARs.

## REFERENCES

- Bain, H. M., Steenburgh, R. A., Onsager, T. G., & Stitely, E. M. 2021, *Space Weather*, 19, e2020SW002670, doi: [10.1029/2020SW002670](https://doi.org/10.1029/2020SW002670)
- Bronarska, K., & Michalek, G. 2017, *Advances in Space Research*, 59, 384, doi: <https://doi.org/10.1016/j.asr.2016.09.011>
- Cliver, E. W., Mekhaldi, F., & Muscheler, R. 2020, *ApJL*, 900, L11, doi: [10.3847/2041-8213/abad44](https://doi.org/10.3847/2041-8213/abad44)
- Cortie, A. L. 1901, *ApJ*, 13, 260, doi: [10.1086/140816](https://doi.org/10.1086/140816)
- Goodfellow, I., Bengio, Y., & Courville, A. 2016, *Deep Learning* (MIT Press)
- Hale, G. E., Ellerman, F., Nicholson, S. B., & Joy, A. H. 1919, *ApJ*, 49, 153, doi: [10.1086/142452](https://doi.org/10.1086/142452)
- He, Y., Yang, Y., Bai, X., et al. 2021, *Advances in Astronomy*, 2021, 5529383, doi: [10.1155/2021/5529383](https://doi.org/10.1155/2021/5529383)
- Jaeggli, S. A., & Norton, A. A. 2016, *ApJL*, 820, L11, doi: [10.3847/2041-8205/820/1/L11](https://doi.org/10.3847/2041-8205/820/1/L11)
- Kasapis, S., Zhao, L., Chen, Y., et al. 2022, *Space Weather*, 20, e2021SW002842, doi: [10.1029/2021SW002842](https://doi.org/10.1029/2021SW002842)
- Kataoka, R., Sato, T., Miyake, S., Shiota, D., & Kubo, Y. 2018, *Space Weather*, 16, 917, doi: [10.1029/2018SW001874](https://doi.org/10.1029/2018SW001874)
- Laitinen, T., Dalla, S., Waterfall, C. O. G., & Hutchinson, A. 2023, arXiv e-prints, arXiv:2303.03168, doi: [10.48550/arXiv.2303.03168](https://doi.org/10.48550/arXiv.2303.03168)
- Martens, P. C. 2018, in *Deep Space Gateway Concept Science Workshop*, Vol. 2063, 3188
- McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. 2016, *SoPh*, 291, 1711, doi: [10.1007/s11207-016-0933-y](https://doi.org/10.1007/s11207-016-0933-y)
- McIntosh, P. S. 1990, *SoPh*, 125, 251, doi: [10.1007/BF00158405](https://doi.org/10.1007/BF00158405)
- Parker, E. N. 1958, *ApJ*, 128, 664, doi: [10.1086/146579](https://doi.org/10.1086/146579)
- Reames, D. V. 2021, *Solar Energetic Particles. A Modern Primer on Understanding Sources, Acceleration and Propagation*, Vol. 978, doi: [10.1007/978-3-030-66402-2](https://doi.org/10.1007/978-3-030-66402-2)
- Sadykov, V., Kosovichev, A., Kitiashvili, I., et al. 2021, arXiv e-prints, arXiv:2107.03911, doi: [10.48550/arXiv.2107.03911](https://doi.org/10.48550/arXiv.2107.03911)
- Toriumi, S., Schrijver, C. J., Harra, L. K., Hudson, H., & Nagashima, K. 2017, *ApJ*, 834, 56, doi: [10.3847/1538-4357/834/1/56](https://doi.org/10.3847/1538-4357/834/1/56)
- Torres, J., Zhao, L., Chan, P. K., & Zhang, M. 2022, *Space Weather*, 20, e2021SW002797, doi: [10.1029/2021SW002797](https://doi.org/10.1029/2021SW002797)
- Whitman, K., Egeland, R., Richardson, I. G., et al. 2022, *Advances in Space Research*, doi: <https://doi.org/10.1016/j.asr.2022.08.006>