The Astrophysical Journal, 979:63 (9pp), 2025 January 20 https://doi.org/10.3847/1538-4357/ad9b27

Identifying Coronal Mass Ejection Active Region Sources: An Automated Approach

Julio Hernandez Camero[1], Lucie M. Green[1], and Alex Piñel Neparidze[2]
[1] Mullard Space Science Laboratory, University College London, UK
[2] University College London, UK
Received 2024 October 2; revised 2024 December 4; accepted 2024 December 4; published 2025 January 17

Abstract

Identifying the source regions of coronal mass ejections (CMEs) is crucial for understanding their origins and improving space weather forecasting. We present an automated algorithm for matching CMEs detected by the Large Angle Spectrometric Coronagraph with their source active regions, specifically Space Weather HMI Active Region Patches (SHARPs), between 2010 May and 2019 January. Our method uses posteruptive signatures, including flares and coronal dimmings, to associate CMEs with potential source regions. Out of 4190 CMEs, we successfully match 1132, achieving a recall rate of ~57% for frontside events. We find that the algorithm performs better for complex SHARP regions containing multiple NOAA regions and for faster CMEs, consistent with expectations that more energetic events produce stronger eruption signatures. We find that CME–flare association rates increase with flare intensity, aligning with previous studies. While our approach has limitations, such as focusing exclusively on SHARP regions and relying on a limited set of posteruptive signatures, it significantly reduces the time required for CME source identification while providing transparent, reproducible results. We encourage the solar physics community to build upon this work, developing improved automated tools for CME source identification. The resulting catalog of CME–source region associations is made publicly available, offering a valuable resource for statistical studies and machine learning applications in solar physics and space weather forecasting.

Unified Astronomy Thesaurus concepts: Solar coronal mass ejections (310); Solar physics (1476); Catalogs (205); Solar flares (1496); Solar active region magnetic fields (1975); Solar active regions (1974); Solar activity (1475)

1. Introduction

Coronal mass ejections (CMEs) are magnetized plasma eruptions from the solar atmosphere and one of the most energetic types of events produced by the Sun. These eruptions inject on the order of 10^{32} erg of energy, 10^{23} Mx of magnetic flux, and 10^{16} g of plasma into the interplanetary medium (T. G. Forbes 2000). When they come in contact with the magnetosphere, the ensuing geomagnetic storms can cause disruption to Earth's technological systems, including satellite operations, communication, and navigation systems or electrical distribution systems (T. Pulkkinen 2007). As such, these events are of great interest not only from a physical perspective, helping us better understand the magnetic mechanisms driving solar activity, but also from a forecasting perspective in order to prevent or minimize their impacts on our technological infrastructure.

In order to understand the physical mechanisms involved in CME occurrence, CME source regions on the Sun must be studied. Individual CME source regions can be investigated through case studies, but statistical and machine learning approaches require large data sets with confident associations of the CMEs to their source regions. This identification is challenging due to the nature of CME detection and the large number of CMEs observed in the current age, where constant observations of the solar corona by coronagraphs mean that more than 2000 CMEs may be detected every year during solar maximum.

CMEs are detected in coronagraphs, such as the Large Angle and Spectrometric Coronagraph (LASCO; G. E. Brueckner et al. 1995) on board the Solar and Heliospheric Observatory (SOHO), which occult the solar disk. This is because CME detections rely on the Thomson scattering of solar photons, which is too faint to be seen unless the disk is occulted. Hence, we cannot directly observe the CME in disk images. Moreover, it is not possible to distinguish between CMEs produced by frontside (Earth-facing) versus backside regions using LASCO. Therefore, we need to rely on signatures left by the CME on the disk to identify the source—though some CMEs, known as stealth CMEs, do not leave any signatures (E. Robbrecht et al. 2009) or the signatures are difficult to detect (N. Alzate & H. Morgan 2017; J. OKane et al. 2019).

There have been several works in the past that use signatures of CME occurrence as observed in the lower solar atmosphere to identify a CME's source region. P. Subramanian & K. P. Dere (2001) used on-disk signatures in images from the EUV Imaging Telescope (EIT) on board SOHO, including erupting prominences, dimmings, and flares, to manually identify the source region of 32 CMEs between 1996 January and 1998 May. They found that CMEs not associated with a prominence eruption typically come from active regions with a lifetime of 11–80 days. Meanwhile, those associated with a prominence eruption come from much older active regions with lifetimes of ~6–7 months. G. Zhou et al. (2003) also used EIT data to find the source regions of 197 frontside halo CMEs by finding flares and filament eruptions that occur within ±30 minutes of the estimated CME onset and where the source region position angle falls within the span of the CME. They find that 79% of the matched CMEs originate from an active region and that all CMEs have an associated brightening in EIT images as well as in Hα. S. Yashiro et al. (2005) studied the

association of CMEs with 1301 flares and the visibility of the CMEs as a function of the flare location and intensity. They found that the rate of association of flares with CMEs increases with flare class, reaching up to 100% for flares above X3 GOES class. S. K. Tiwari et al. (2015) manually identified the source region of 189 CMEs and demonstrated that the speed of the fastest CME an active region can produce correlates with free-magnetic-energy proxies and magnetic twist parameters from vector magnetograms. S. A. Murray et al. (2018) used ballistic back-propagation of CMEs observed by the Solar TErrestrial RElations Observatory spacecraft to automatically match them to flares and their source regions. The 550 matched CMEs are released as the HELCATS LOWCAT catalog. In a more recent work, S. Majumdar et al. (2023) manually matched 3327 CMEs from 1998 to 2017.

In this work, given a CME detected by LASCO, we find its source active region using posteruptive signatures that can be detected on the disk. We focus only on CME source regions that are located within Space Weather HMI Active Region Patches (SHARPs; M. G. Bobra et al. 2014). Our motivation for using an automated approach is the significantly reduced human time required to produce these matches and removing the susceptibility of the associations to changes in personnel or their level of experience. We are also driven by the increased transparency in the associations that an automated approach can provide; i.e., the reasoning for each association is clearly and simply described, which is often lacking in human-made associations, especially if an end user is interested in quickly filtering these associations based on their confidence in the association methods. Moreover, even though any automated algorithm can be prone to errors, we hope that by providing the algorithm as an open-source project, these can be characterized easily and regularly through smaller, curated verification data sets leading to community-driven improvements to the algorithm. However, we point out that we release this work as a proof of concept, showing how an automated approach to this task is possible, and not as an operational tool.

In Section 2, we describe the posteruptive signatures used in our algorithm, in Section 3 we describe the data used in this work and in Section 4, we describe how they are used to make the associations. Section 5 describes the procedure used to manually check a number of associations made by our method. In Section 6, we show a small sample of summary statistics of the resulting associations, and we discuss their implications in Section 7. We conclude and discuss future developments of this work in Section 8.

2. Posteruptive Signatures

With posteruptive signatures, we refer to signatures left by the CME that manifest themselves in the corona or the chromosphere. This can include the following.

1. Dimmings—a reduction of the brightness in the lower corona, usually in EUV bands (E. Kraaikamp & C. Verbeeck 2015)—have been associated with CMEs and are likely caused by mass evacuation during the eruption. Mass-loss calculations using dimmings have been matched to mass estimations from CME observations to support the mass-loss theory (R. A. Harrison et al. 2003). The relationship between CMEs and dimmings has been statistically proven by D. Bewsher et al. (2008) and studied by K. Dissauer et al. (2019), which makes them extremely useful in tracing a CME back to its source region.
2. Flares are also known to be related to CMEs (B. C. Low 1996; T. G. Forbes 2000; Z. Švestka 2001; B. Vršnak 2008). However, not all CMEs seem to have an associated flare, nor do all flares have an associated CME. Flares that do not have an associated CME are known as "confined flares," and it is possible that the reason some CMEs do not have an associated flare is that the flare is too weak to be detected.
3. EUV waves—large-scale disturbances that propagate through the Sun's atmosphere in the form of increased EUV emission (E. Kraaikamp & C. Verbeeck 2015)—have also been shown to be linked to CMEs (S. Patsourakos & A. Vourlidas 2009; S. Patsourakos et al. 2009; P. T. Gallagher & D. M. Long 2011).

Given the connection of these signatures to CME eruptions, we may use them to identify the source of a CME. For this work, flares and dimmings are used, since catalogs are available as described in Section 3.

3. Data Sources

Our study makes use of various existing catalogs and data sources, which are described here.

1. *SHARP active region bounding boxes*. We make use of the bounding boxes defined for SHARP active regions. These regions are defined by tracking coherent magnetic structures in line-of-sight data from the Helioseismic and Magnetic Imager (HMI; P. H. Scherrer et al. 2012) on board the Solar Dynamics Observatory (SDO) as they rotate following the Sun's rotation. SHARP active regions do not need to have a sunspot, and while they are often associated with NOAA active regions, they do not necessarily have to and may even be associated with more than one NOAA active region. Thus, SHARP regions cover both active regions with strong magnetic fields and more dispersed, aged active regions. These bounding boxes are crucial information about the position and spatial extent of the regions that we use in our work. We obtain the SHARP bounding boxes from R. A. Angryk et al. (2020). Although they could be requested directly from the Joint Science Operations Center, using the preprocessed data by R. A. Angryk et al. (2020) was found to be more efficient for our work. We included an extra preprocess step as we found some oversized bounding boxes encompassing a disproportionate area of the solar disk, along with smaller regions that substantially overlap with larger ones in both spatial extent and temporal duration. These regions are removed, and the details can be found in the source code.
2. *CMEs*. We use the CDAW CME catalog[3] from the years 2010–2018. This catalog is created manually with data from the C2 and C3 coronagraphs. For CMEs with an unclear identification, a "Poor Event" or "Very Poor Event" comment is added. We discount these events to use only confident CME identifications in this study. The period we cover contains 4190 such CMEs.
3. *Flares*. We make use of the curated flare list from R. A. Angryk et al. (2020), in which the locations and association between flares and SHARP regions are

[3] This CME catalog is generated and maintained at the CDAW Data Center by NASA and the Catholic University of America in cooperation with the Naval Research Laboratory. SOHO is a project of international cooperation between ESA and NASA. https://cdaw.gsfc.nasa.gov/CME_list/.

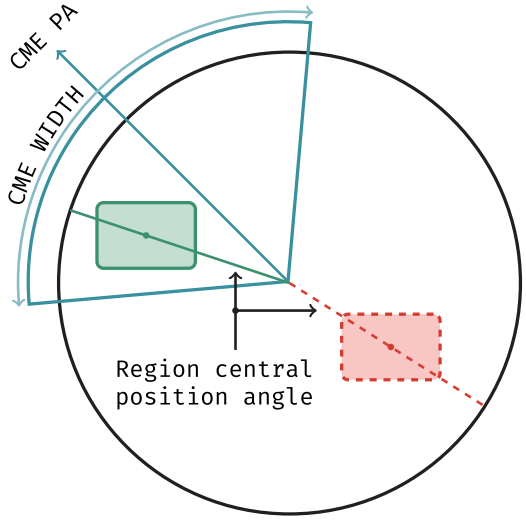

Figure 1. Possible source regions for a CME are selected by considering its width and position angle. Any region within a wedge centered on the CME's position angle and width equal to the CME's width is considered. Here, the green rectangle represents a region that would be considered as a potential source region for the CME. Meanwhile, a region represented by the dashed rectangle would not be a potential source.

verified by comparing and aggregating flare detections from GOES (SWPC 2019), SolarSoftWare (SSW) latest events (S. Freeland 2018), and Hinode X-Ray Telescope (XRT) (K. Watanabe et al. 2012). The data from R. A. Angryk et al. (2020) cover the 2010–2018 range and are the factor limiting the time range considered in our work.

4. *Dimmings*. Dimmings are obtained from Solar Demon (E. Kraaikamp & C. Verbeeck 2015). This is an automated tool for the near-real-time detection of dimmings from 211 Å images from the Atmospheric Imaging Assembly (AIA; J. R. Lemen et al. 2012) on board SDO. For each dimming, a peak time and position (from the intensity-weighted mass center) are determined.

4. The Automated Algorithm

For each CME included in our list, the algorithm executes the following steps.

1. Finds the SHARP region(s) that have a central position angle within the CME's span as seen in LASCO data. See Figure 1 for an example.
2. Identifies which of the SHARP regions from the previous step had signatures that are potentially indicative of a CME occurrence.
3. Ranks the potential source regions based on these signatures to choose the best-ranking one as the source.

This sequence of steps is summarized in Figures 2 and 3. In the case when no CME source region is identified, either because there were no SHARP regions under the span of the LASCO CME or no flare/dimming was observed, the CME is left without an associated source region.

While flares have been associated with their source SHARP region by R. A. Angryk et al. (2020), Solar Demon does not provide such associations. Therefore, their source SHARP region must be found using the intensity-weighted barycenter at the dimming's peak extent as its position. For each dimming, the closest region that is less than 10° from it—with the distance measured from the dimming location to the edges of the region's bounding box or 0 if the dimming is inside the bounding box—is matched to the dimming. If there is no region within 10°, the dimming is unmatched.

The result of steps 1 and 2 will be, for each CME, a list of regions that were at the right location to be the source of this CME. If one of these regions has produced one of the signatures recently, e.g., a flare, we require that it has happened within 2 hr before the CME was detected in LASCO in order to consider it in step 3. This is solely based on the possible travel time of the CME from the disk to LASCO'S field of view.

Given the conditions above, it is possible that more than one SHARP region is a candidate for a CME source (e.g., two regions are under the span of the CME, and both had a flare within 2 hr before the CME was detected in LASCO). To handle these situations, we create a series of scores roughly indicating our confidence in the match. These scores are based on the signatures that each region exhibits and are detailed in the right panel of Figure 2. Once each region is assigned a score, the one with the lowest number is assigned as the source of the CME.

Running the whole script for the years 2010–2018, which includes more than 4000 CMEs, takes about 14 minutes on a consumer laptop. However, if a similar script to this were implemented in an operational setting, the processing of CMEs could be performed in near-real time (as data become available), since the time to process a single event is expected to be very short.

5. Manual CME Source Active Region Association Verification Procedure

In order to identify possible errors in the algorithm, we perform a manual verification of the associations. This allows us to quantify the accuracy of our automated approach and characterize any errors. Given the extent of the data set, we have selected a sample set of 300 randomly selected associations for manual verification.

Our manual verification methodology consists of comparing available visual data for each CME signature used in the association process. We obtain these data by using the unique flare and dimming identifiers provided for each association, which we then use to gather AIA 211 Å atmospheric images for flare events and dimming masks images from Solar Demon. By comparing the timing of these events, as well as their coordinates on the solar surface, we manually determine if their association with the CME and SHARP region is correct.

The multitude of data required for manual verification makes the process inefficient and time-consuming. To address this problem, we have developed CatView, an open-source application that automatically gathers all the data needed for this process into a comprehensive workspace. This substantially reduced the time needed to validate each CME source region. The application allows the user to parse through the association catalog, automatically showing the necessary Solar Demon and SDO data for each event, alongside a map of SHARP bounding boxes and other data used for the automatic association.

Throughout the verification process, we use strict criteria when labeling associations as correct or otherwise, only accepting those that clearly showcase matching signatures as correct. In cases where the data are ambiguous or difficult to interpret, we have decided to label the association as incorrect while also providing a brief explanation for future reference.

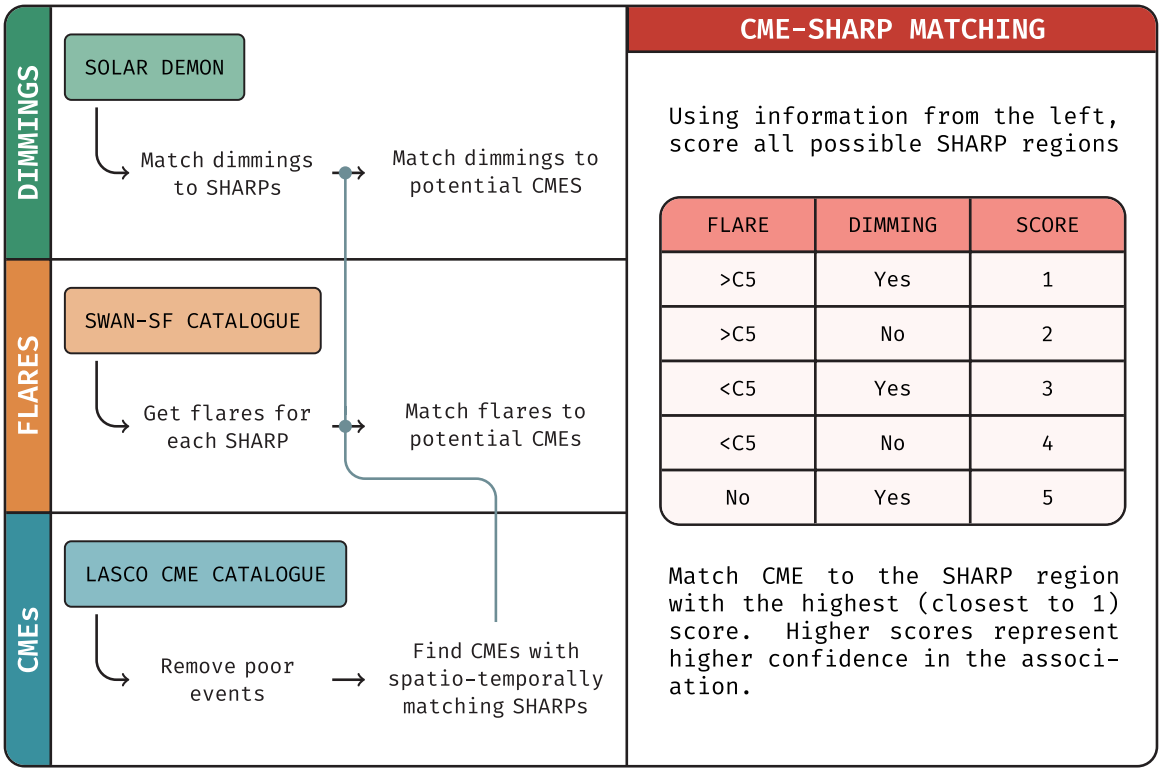

Figure 2. Simplified process for the creation of the CME source region catalog. We make use of posteruptive signatures in the form of dimmings from Solar Demon (E. Kraaikamp & C. Verbeeck 2015) and flares from the SWAN-SF catalog (R. A. Angryk et al. 2020) and consider CMEs from the LASCO CDAW CME catalog. The vertical line crossing from CMEs to flares and dimmings indicates that we consider the dimmings and flares associated with SHARP regions that are potential sources for a particular CME (see Figure 1). Each potential SHARP region is given a score according to the right panel. If more than one region is a possible source, the region with the lowest score is taken.

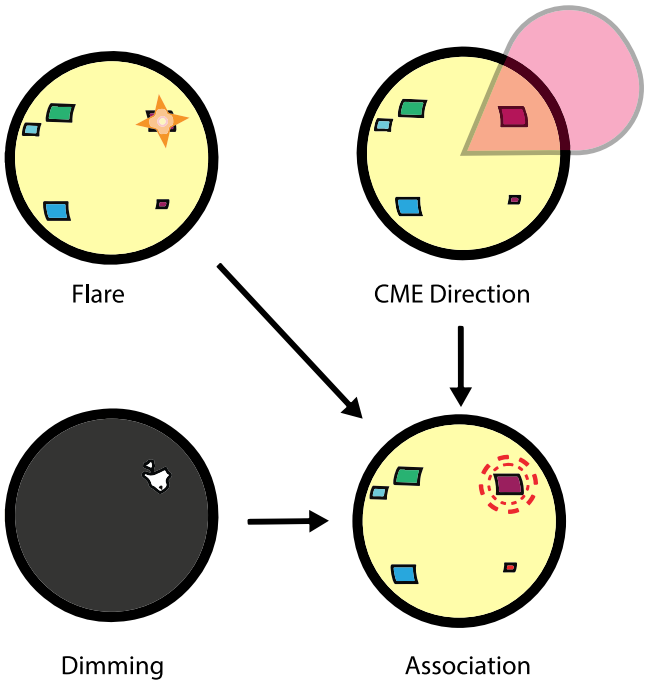

Figure 3. The association of a CME with a source SHARP region (bottom right; colored boxes in this figure represent the bounding boxes of different regions) depends on whether the span of the CME (top right) covers any SHARP region and on the region producing at least a flare (top left) or a dimming (bottom left, showing the dimming mask) associated with this CME.

While we find this manual approach to be largely effective at validating events, certain challenges and issues have been identified. Associating CMEs with individual active regions, and particularly with a single SHARP bounding box, has been found to be a limitation. Throughout the manual verification process, we have found multiple instances where CME posteruptive signatures extend across multiple SHARP regions and cannot be associated with a single one.

We also find the use of AIA atmospheric images to be particularly limiting, especially for weak <C5 flares; the faintness of these events makes it difficult to clearly discern their source, potentially leading to a negative bias toward associations connected to such flares.

6. Results

We match 1132 CMEs out of 4190 recorded by LASCO CDAW with no quality flags during the time period of this study. This means we recover about 27% of the CMEs, when we would expect to recover around 50% when accounting for the fact that source regions for CMEs in the farside of the Sun will not be visible with the data used in this work. Therefore, our algorithm has a recall of ~57%. Given that we only match CMEs originating from active regions, when we compare this to G. Zhou et al. (2003), who found that 79% of their matched CMEs come from active regions, this means we probably are not recovering all active region CMEs from the frontside of the Sun. Possible reasons for the relatively small recovery rate will be discussed in Section 7. In Figure 4, we show how these matched CMEs are distributed by verification score. The most

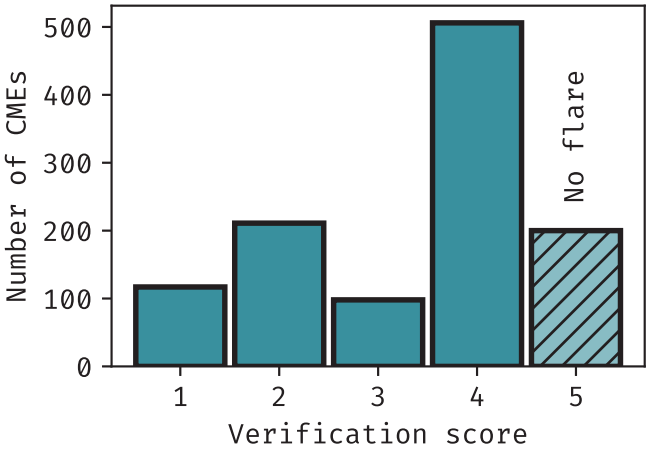

Figure 4. Distribution of verification scores for the matched CMEs. Score 5 is highlighted as it corresponds to CMEs without an associated flare.

common (~500 matches, compared to the second-most-common, ~200) is score 4, corresponding to a match made only with a flare of class less than C5. Since small flares are common, this is to be expected. At any given point, it will be likely to find a small flare that could match with a CME, and so this score is also the least reliable. Hence, we say these verification scores represent confidence in the match loosely.

In Figure 5, we show the flare–CME association rate for different flare class bins in comparison with the results of S. Yashiro et al. (2005). While our overall flare–CME association rates are lower, we observe the same trend of increasing association rates with higher flare classes. The lower association rates are likely due to our limited recovery rate and the small sample size in the higher flare class bins.

In order to further characterize the behavior of our algorithm, i.e., which types of CMEs and source regions are being matched, biases, etc., we plot the following statistics.

First, we consider what kind of SHARP regions are being matched. Since each SHARP region may contain zero, one, or more than one NOAA active region, we plot the distribution of the number of NOAA active regions per SHARP region. We do this for both the whole population of SHARP regions in the 2010–2018 range and those that were matched to a CME in Figure 6. We find that the more NOAA active regions there are in a SHARP region, the more likely it is to have been associated with a CME.

To investigate how this may be related to the signatures our algorithm is considering, we show in Figure 7 how this distribution changes per verification score. Note here that a comparison with the full population is not possible, as a match is needed to assign a verification score. We find that the bulk of SHARP regions with zero active regions, expected to correspond to a more dispersed field without sunspots, are matched with score 5. This corresponds to a dimming with no flare. Given that these SHARP regions with no NOAA active regions within them should correspond to a more dispersed magnetic field without a sunspot, they may have insufficient free magnetic energy to produce a detectable flare. So having most of the matches originating from a SHARP region with no NOAA active regions in verification level 5 is expected.

Focusing on the distribution of CME speeds of matched CMEs compared to that of the whole population, we plot the distribution of speeds of all CMEs versus the distribution of the matched CMEs. This comparison is shown in Figure 8. The

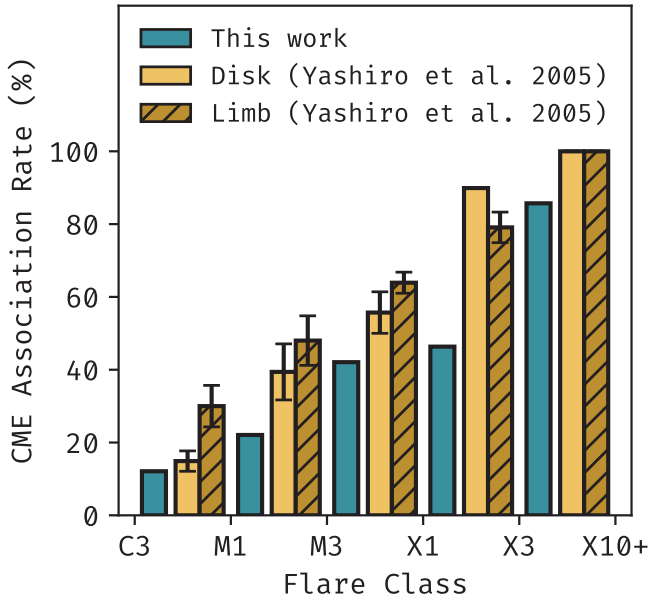

Figure 5. Flare association rates with CMEs. Result from this work are compared with S. Yashiro et al. (2005).

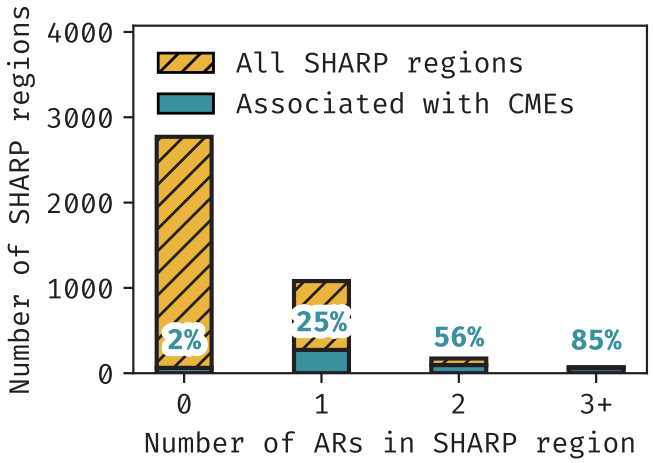

Figure 6. Distribution of number of NOAA active regions per SHARP region for all SHARP regions and those that were matched to a CME. Percentages show, for each number of active regions, how many were matched to a CME. For an unbiased algorithm, we would expect these percentages to be constant across the *x*-axis. The increasing percentages instead indicate a bias in the algorithm to detecting CMEs from regions that contain a larger number of active regions.

larger the speed of the CME, the more likely it is to be matched to a source region by our algorithm. These more energetic CMEs can be expected to originate from equally more energetic and complex source regions that are more likely to produce the signatures we are looking for clearly. Hence, our algorithm seems to be biased toward large and complex regions, which produce fast CMEs.

Finally, considering the distribution of the Stonyhurst longitudes of the source regions in Figure 9, we find a dip in regions with latitudes within 15° of Sun center in nonhalo CMEs.

6.1. Evaluation of Catalog Accuracy

Figure 10 shows the error rates per verification score from our manual verification. We have found that associations with verification scores 3 (dimming and flare <C5) and 5 (dimming

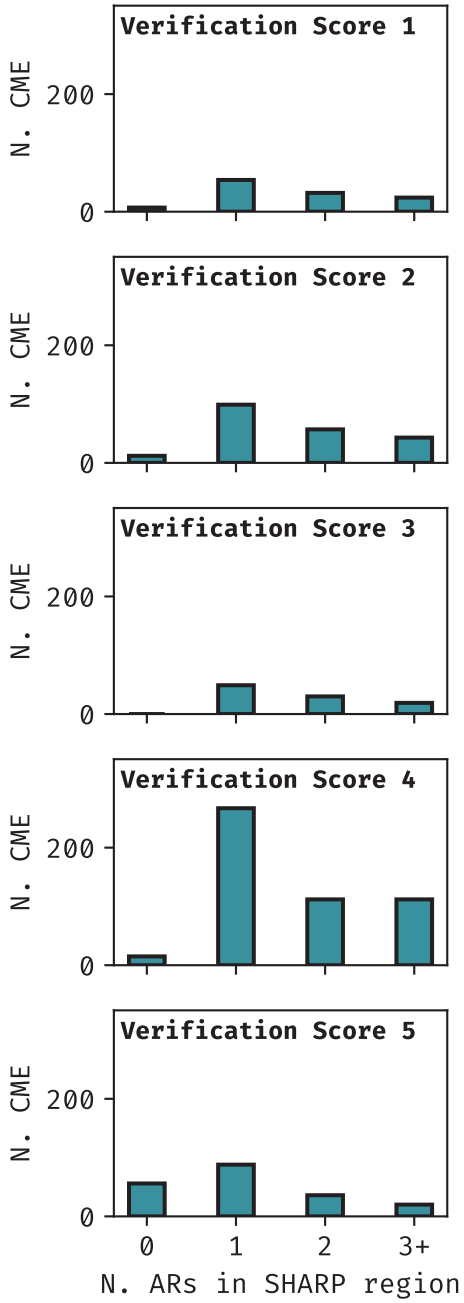

Figure 7. Distribution of number of NOAA active regions within each SHARP for CMEs by verification score from the right panel of Figure 2.

with no flare) are less error-prone than anticipated, with error rates of 3.23% and 31.3%, respectively. This is particularly noteworthy given that events with scores 2 (flare >C5 and no dimming) and 4 (flare <C5 and no dimming) have substantially higher error rates than expected at 28.6% and 56.4%, respectively. Hence, we say that the verification scores only describe the confidence in the association loosely.

Given that scores 3 and 5 include dimmings as part of their associations and that scores 2 and 4 do not, this appears to indicate that dimmings serve as a more reliable CME signature than flares alone, at least given our matching procedure. For example, comparing stronger flares (greater than C5) with and without dimmings shows that including dimmings improves the accuracy of the match. The same trend holds for weaker flares (less than C5) when comparing them with and without

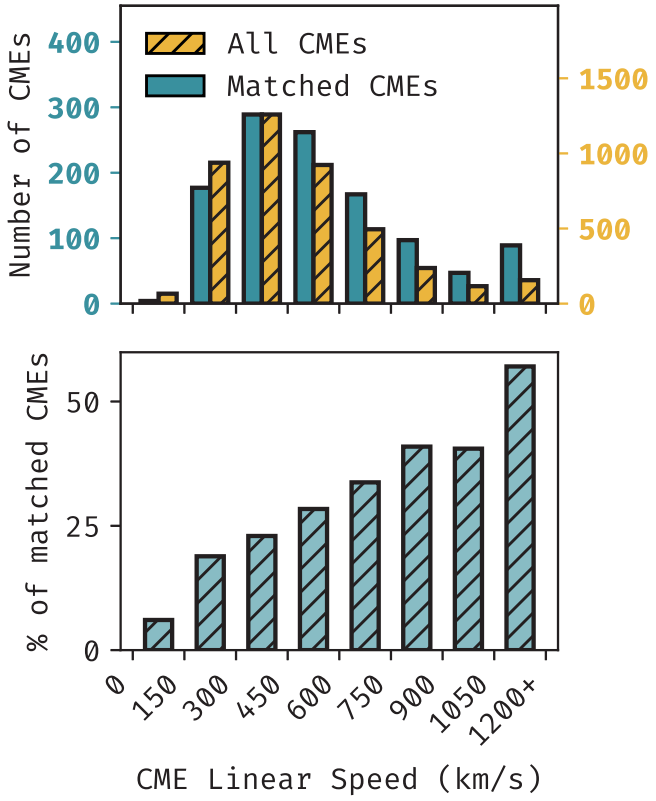

Figure 8. Distribution of CME speeds for all CMEs and those that are matched by our algorithm. The top panel shows the distribution with two different *y*-axis scales. For matched CMEs, the distribution is biased to slightly larger speeds compared to all CMEs. In the bottom panel, we show for each of the bins in the top panel what percentage of all CMEs have been matched, again showing a bias toward faster CMEs in our algorithm.

dimmings. Furthermore, when comparing weaker flares alone to dimmings alone, dimmings are a more reliable indicator of CME eruptions than weaker flares. The better accuracy in the associations could be due to either a much higher number of weak flares, facilitating false associations, or dimmings being a better indicator of a CME eruption.

7. Discussion

In Figure 6, we find that SHARP regions with more NOAA active regions within them are more likely to be matched to a CME. We expect regions containing more NOAA active regions to be larger and thus contain more magnetic energy. These regions are also more likely to contain complex magnetic configurations with multiple polarity inversion lines. This means that they are more likely both to be CME productive and to produce a flare or a dimming when they do erupt. There are then two contributions to the trend seen in Figure 6.

1. We expect SHARP regions with fewer NOAA active regions—especially those without any, corresponding to a dispersed field—to be less CME productive. On the other side, we expect SHARP regions with many NOAA active regions to be energetic and complex, making them more likely to produce a CME.
2. Even if the CME productivity was equal regardless of the number of NOAA active regions, more complex and energetic regions are more likely to produce the signatures that we are looking for, especially flares, compared to dispersed field regions.

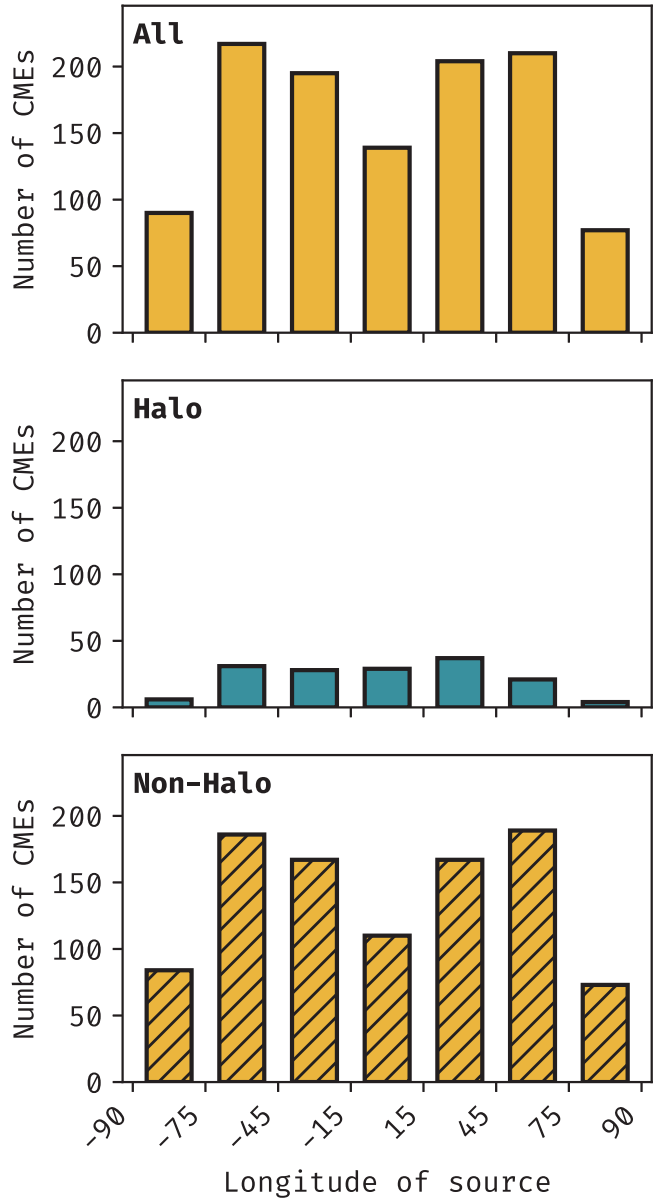

Figure 9. Distribution of Stonyhurst longitudes for CME source regions. There is a dip for regions within 15° of central longitude, probably due to CMEs originating from these regions being harder to detect in LASCO. There are also dips at either extreme of longitude; however, also note that the bin size for the extremes is half the size of the others.

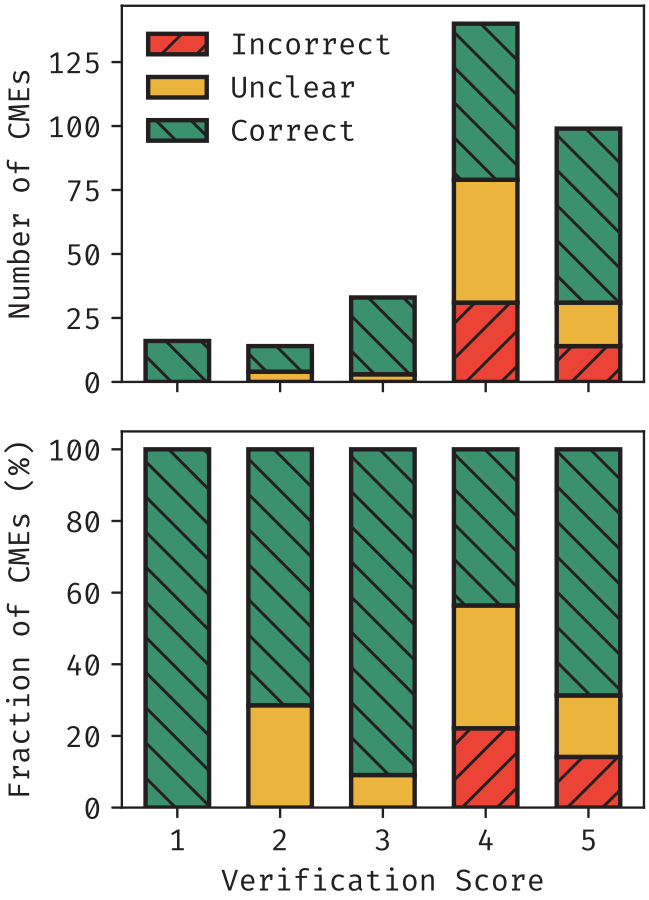

Figure 10. Results of the manual verification of 300 randomly selected events from the catalog. We show the number of incorrect, unclear (where, e.g., the CME source involved an interaction between two regions), and correct associations per verification level and also the normalized values.

To what extent the trend in Figure 6 is due to a bias in our algorithm caused by the signatures we are looking for (second contribution) and not just due to larger regions being more likely to produce CMEs (first contribution) is hard to say. To shed light on this, we would need a catalog that is known to have matched all the frontside CMEs for the same time period and compare the distributions, which at this time is not available.

Regarding the dip in the longitude distribution of Figure 9, a similar gap was found by S. Majumdar et al. (2023; see their Figure 16). This is likely due to the decreased Thomson scattering efficiency for CMEs originating from central regions making them harder to detect in white light (A. Vourlidas & R. A. Howard 2006), meaning they are less frequent in the LASCO CDAW catalog, rather than a bias in the algorithm. This is supported by the fact that dimmings are more frequently detected in central longitudes, and flares do not show any dips at those longitudes as shown in Figure 11, meaning we do not expect these longitudes to represent a special challenge for the algorithm when matching CMEs.

The relatively small number of recovered CMEs can be attributed to several factors.

1. *Exclusive focus on SHARP regions*. This method only considers CMEs with source regions that are located within SHARPs. Consequently, CME source regions that are located outside the SHARP regions will not be detected. For example, CMEs that originate from polarity inversion lines in the very weak and dispersed field at high latitude are associated with the eruption of polar crown filaments. Incorporating a catalog of filaments into our algorithm would allow us to check for the disappearance of a filament at the time of the CME and link the two events together. However, it would require a completely different approach to identifying the source region, as there may not be a SHARP number and bounding box to assign as the source.
2. *Limited set of signatures*. The algorithm relies on flares and dimmings as signatures. CMEs that do not produce these particular signatures remain undetected. For example, CMEs from weak field regions will produce weak flare emissions that may not be detected against background radiation levels so that no flare will be recorded.

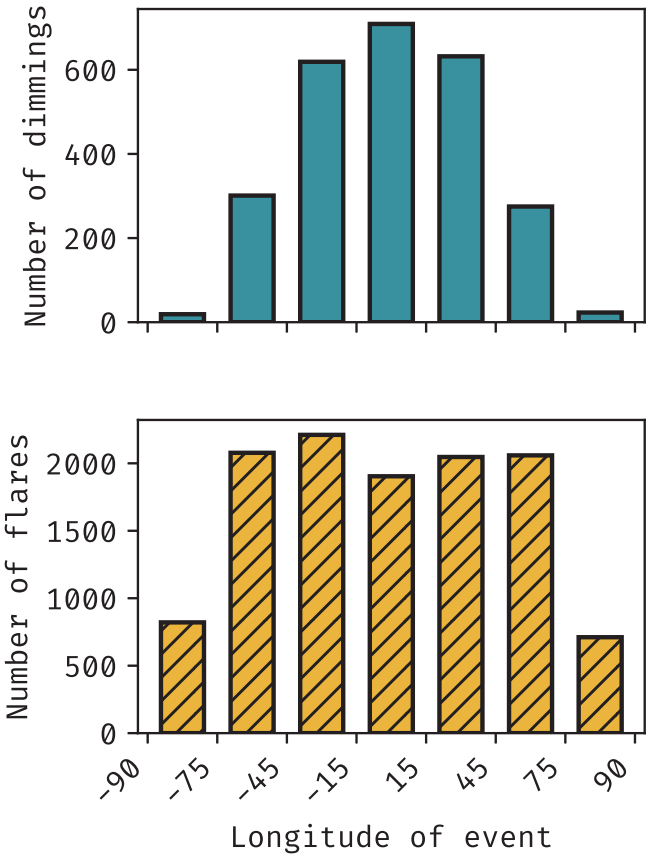

Figure 11. Longitude distribution for all flares and dimmings.

3. *Association of dimmings with SHARPs*. Our method of association of dimmings with SHARP regions relies on defining the position of the dimming as a single point. Dimmings have significant spatial extents, and this method could lead to associating the dimming with the wrong region or missing the association altogether, which would negatively impact the association of CMEs. During the time period of our study, there were 4740 dimmings detected by Solar Demon. We matched 2579; therefore, many dimmings were missed, either because they did not originate within a SHARP region or because relying on a single point to define their position led to our algorithm missing their source. Some of these missing dimmings could even be associated with a filament eruption.

However, these are all problems that could be solved by future iterations of the algorithm. The success of this proof-of-concept implementation at finding more than 1000 CME source regions highlights the potential of this approach to become a regular tool in the solar and forecaster community.

While the manual verification process proved that the algorithm is highly reliable in its performance, we have also identified certain errors and misclassifications in the resulting catalog. However, these errors appear to be caused under certain conditions, mainly the types of signatures used to perform the associations between active regions and CMEs, as was observed in the high error rate of associations at levels 4 and 5. At the same time, the identified errors are also largely consistent with the limitations that arise from using SHARP regions and their absolute separation of active regions that could, in fact, be connected.

It must also be noted that the algorithm may be prone to perform some misassociations by pure chance. A random signature event may occur at the same time as a CME and be incorrectly associated with an active region that did not cause the eruption. This is also a likely explanation as to why associations performed with only weak <C5 flares perform worse than those that only used dimmings; weak flares are a far more common occurrence and are therefore more likely to trigger a misassociation.

Finally, the relatively small recovery rate of our algorithm means that there will be flares that were associated with a CME but were not matched in our catalog. This could explain why in Figure 5, the association rates are smaller than those of S. Yashiro et al. (2005). However, the trend that higher flare classes are associated with higher CME association rates is reproduced in our results. It is also worth noting that although S. Yashiro et al. (2005) report a 100% association rate for flares larger than X3, they point out that not all flares of this type have an associated CME. The number of flares in this bin is very small, seven in our data with six having a CME, which means it can suffer from small number statistics problems.

8. Conclusion

We have developed an automated algorithm for the identification of the active region sources of CMEs using posteruptive signatures. Our algorithm successfully matched 1132 CMEs with their source regions in the period from 2010 to 2018, demonstrating the potential for automation of this task. We provide the list of the matched CMEs for use by the general scientific community. This data set may be used for diverse tasks, such as training machine learning models for CME forecasting and for statistical studies on the pre- and post-CME evolution of source regions. An automated approach to this task significantly reduces the need for manual intervention, saving valuable researcher time while providing a fast way to identify the source region for CMEs. Our approach is also fully transparent, with the source code publicly available. Each association follows a clear logic that can be scrutinized and improved by the community. This kind of transparency can increase confidence in the associations and drive consensus on data sets for tasks such as machine learning forecasting of CMEs.

Although automated approaches such as ours are promising, we acknowledge their limitations, including a relatively poor recovery rate due to design constraints. Most importantly, our work is limited to SHARP regions—meaning any CME not from a SHARP region will be missed—and relies on a specific set of signatures that are not always present after a CME. Improvements should focus on extending the algorithm to other source region types and signatures. For example, jets have been associated with stealth CME sources (N. Alzate & H. Morgan 2017), and image processing techniques can improve the detection of posteruptive signatures.

We also note that every mistake made in the algorithm design has the potential to propagate to all associations. While this may make it easier to detect errors, undetected errors would have a larger impact than in human-made associations. A possible solution, which we follow here, is to host the algorithm as an open-source project. In this way, regular inspections offer the possibility of identifying such errors as well as performance degradation due to changes in any of the

data sources. However, this would require regular input from field experts and their time investment.

To make the improvement of automatic CME source detection algorithms more transparent, we consider the creation of a citizen science project that would result in a curated verification catalog for benchmarking future iterations of automated algorithms. With a large enough curated data set, machine learning approaches would also become possible. We believe that there is potential in the integration of future iterations of this algorithm with current tools in space weather. The clear identification of numerous source regions of CMEs would provide a much-needed resource in the investigation of CMEs. We hope to further encourage other automated tools for detecting source regions of CMEs, leading to better catalogs.

While our automated approach removes the time barrier to the identification of large numbers of CME source regions, it also underlines that in order to keep improving our algorithms, expert input remains crucial. Future work will make use of the generated list by training a machine learning model to produce forecasts of CMEs and a statistical study of source regions with respect to their evolution pre- and post-CME.

Data Availability

The source code for the algorithm for matching CMEs with source regions and to reproduce the figures in this work is made available through a Zenodo upload (J. Hernandez Camero 2024). The data required to run the algorithm as well as the final version of our catalog are also available in a Zenodo upload via Doi:10.5281/zenodo.13150638.

The CatView visualization tool is available at A. Pinel Neparidze (2024).

Acknowledgments

This research used version 6.0.0 (S. J. Mumford et al. 2024) of the SunPy open-source software package (The SunPy Community et al. 2020).

J.H.C. acknowledges the support of the Science and Technology Facilities Council (STFC) under grant number ST/X508858/1 for this work.

Facilities: SDO (HMI), SDO (AIA), SOHO (LASCO), GOES.

Software: Astropy (Astropy Collaboration et al. 2013, 2018, 2022), Numpy (C. R. Harris et al. 2020), Pandas (W. McKinney 2010; pandas development team 2020), SunPy (The SunPy Community et al. 2020; S. J. Mumford et al. 2024), Matplotlib (J. D. Hunter 2007).

ORCID iDs

Julio Hernandez Camero https://orcid.org/0000-0002-4472-4559
Lucie M. Green https://orcid.org/0000-0002-0053-4876
Alex Piñel Neparidze https://orcid.org/0009-0009-0001-1360

References

Alzate, N., & Morgan, H. 2017, ApJ, 840, 103
Angryk, R. A., Martens, P. C., Aydin, B., et al. 2020, NatSD, 7, 227
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Bewsher, D., Harrison, R. A., & Brown, D. S. 2008, A&A, 478, 897
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Brueckner, G. E., Howard, R. A., Koomen, M. J., et al. 1995, SoPh, 162, 357
Dissauer, K., Veronig, A. M., Temmer, M., & Podladchikova, T. 2019, ApJ, 874, 123
Forbes, T. G. 2000, JGRA, 105, 23153
Freeland, S. 2018, SolarSoft Latest Events, https://www.lmsal.com/solarsoft/latest_events/
Gallagher, P. T., & Long, D. M. 2011, SSRv, 158, 365
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Harrison, R. A., Bryans, P., Simnett, G. M., & Lyons, M. 2003, A&A, 400, 1071
Hernandez Camero, J. 2024, Identifying Coronal Mass Ejection Active Region Sources: An automated AapproachAutomated Approach - Source code, v0.1.1, doi:10.5281/ ZENODO.14040715
Hunter, J. D. 2007, CSE, 9, 90
Kraaikamp, E., & Verbeeck, C. 2015, JSWSC, 5, A18
Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, SoPh, 275, 17
Low, B. C. 1996, The Role of Coronal Mass Ejections in Solar Activity in Proceedings of the 16th (sixteenth) international workshop National Solar Observatory1995, 95 (Astronomical Society of the Pacific (ASP)), 148
Majumdar, S., Patel, R., Pant, V., et al. 2023, ApJS, 268, 38
McKinney, W. 2010, Data Structures for Statistical Computing in Python in Proceedings of the 9th Python in Science Conference 2010, ed. S. van der Walt & J. Millman, 56
Mumford, S. J., Freij, N., Stansby, D., et al. 2024, sunpy: A Core Package for Solar Physics, v6.0.1, doi:10.5281/zenodo.13743555
Murray, S. A., Guerra, J. A., Zucca, P., et al. 2018, SoPh, 293, 60
OKane, J., Green, L., Long, D. M., & Reid, H. 2019, ApJ, 882, 85
pandas development team, 2020, pandas-dev/pandas: Pandas, latest, doi:10.5281/zenodo.3509134
Patsourakos, S., & Vourlidas, A. 2009, ApJL, 700, L182
Patsourakos, S., Vourlidas, A., Wang, Y. M., Stenborg, G., & Thernisien, A. 2009, SoPh, 259, 49
Pinel Neparidze, A. 2024, Identifying Coronal Mass Ejection Active Region Sources: An Automated Approach - CatView, v0.1.1, doi:10.5281/ZENODO.14185777
Pulkkinen, T. 2007, LRSP, 4, 1
Robbrecht, E., Patsourakos, S., & Vourlidas, A. 2009, ApJ, 701, 283291
Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, SoPh, 275, 207
Subramanian, P., & Dere, K. P. 2001, AAS, 561, 372
Švestka, Z. 2001, SSRv, 95, 135
SWPC 2019, Space Weather Prediction Center (SWPC) Historical SWPC Products and Data Displays, ftp://ftp.swpc.noaa.gov/pub/warehouse (2019)
The SunPy Community, Barnes, W. T., Bobra, M. G., et al. 2020, ApJ, 890, 68
Tiwari, S. K., Falconer, D. A., Moore, R. L., et al. 2015, GeoRL, 42, 5702
Vourlidas, A., & Howard, R. A. 2006, AAS, 642, 1216
VrŠnak, B. 2008, AnGeo, 26, 3089
Watanabe, K., Masuda, S., & Segawa, T. 2012, SoPh, 279, 317
Yashiro, S., Gopalswamy, N., Akiyama, S., Michalek, G., & Howard, R. A. 2005, JGRA, 110, A12
Zhou, G., Wang, J., & Cao, Z. 2003, A&A, 397, 1057