University of Reading

The Solar Stormwatch CME catalogue: results from the first space weather citizen science project

Article

Published Version

Creative Commons: Attribution 3.0 (CC-BY)

Open Access

Barnard, L. ORCID: https://orcid.org/0000-0001-9876-4612, Scott, C. ORCID: https://orcid.org/0000-0001-6411-5649, Owens, M. ORCID: https://orcid.org/0000-0003-2061-2453, Lockwood, M. ORCID: https://orcid.org/0000-0002-7397-2172, Tucker-Hood, K., Thomas, S., Crothers, S., Davies, J. A., Harrison, R., Lintott, C., Simpson, R., O'Donnell, J., Smith, A. M., Waterson, N., Bamford, S., Romeo, F., Kukula, M., Owens, B., Savani, N., Wilkinson, J., Baeten, E., Poeffel, L. and Harder, B. (2014) The Solar Stormwatch CME catalogue: results from the first space weather citizen science project. Space Weather, 12 (12). pp. 657-674. ISSN 1542-7390 doi: 10.1002/2014SW001119 Available at https://centaur.reading.ac.uk/38405/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.
Published version at: http://dx.doi.org/10.1002/2014SW001119
To link to this article DOI: http://dx.doi.org/10.1002/2014SW001119

Publisher: American Geophysical Union

University of Reading

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

AGU PUBLICATIONS

Space Weather

RESEARCH ARTICLE

10.1002/2014SW001119

Key Points:
- Solar Stormwatch has produced a unique CME catalogue, using STEREO/HI images
- The CMEs are tracked over multiple position angles and out to large elongations
- The full data set is publicly available online

Supporting Information:
- Readme
- Movie

Correspondence to:
L. Barnard,
l.a.barnard@reading.ac.uk

Citation:
Barnard, L., et al. (2014), The Solar Stormwatch CME catalogue: Results from the first space weather citizen science project, *Space Weather*, *12*, 657–674, doi:10.1002/2014SW001119.

Received 11 SEP 2014
Accepted 26 OCT 2014
Accepted article online 31 OCT 2014
Published online 1 DEC 2014

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The Solar Stormwatch CME catalogue: Results from the first space weather citizen science project

L. Barnard[1], C. Scott[1], M. Owens[1], M. Lockwood[1], K. Tucker-Hood[1], S. Thomas[1], S. Crothers[2], J. A. Davies[2], R. Harrison[2], C. Lintott[3], R. Simpson[3], J. O'Donnell[3], A. M. Smith[4], N. Waterson[5], S. Bamford[6], F. Romeo[7], M. Kukula[7], B. Owens[7], N. Savani[8], J. Wilkinson[9], E. Baeten[9], L. Poeffel[9], and B. Harder[9]

[1]Department of Meteorology, University of Reading, Reading, UK, [2]RAL Space, Rutherford Appleton Laboratory, Chilton, UK, [3]Astrophysics Department, University of Oxford, Oxford, UK, [4]GitHub Inc, San Francisco, California, USA, [5]National Maritime Museum, Greenwich, UK, [6]Centre for Astronomy and Particle Theory, University of Nottingham, Nottinghamshire, UK, [7]Royal Observatory Greenwich, Royal Museums Greenwich, London, UK, [8]Naval Research Laboratory, Washington, District of Columbia, USA, [9]Zooniverse, c/o Astrophysics Department, University of Oxford, Oxford, UK

Abstract Solar Stormwatch was the first space weather citizen science project, the aim of which is to identify and track coronal mass ejections (CMEs) observed by the Heliospheric Imagers aboard the STEREO satellites. The project has now been running for approximately 4 years, with input from >16, 000 citizen scientists, resulting in a data set of >38, 000 time-elongation profiles of CME trajectories, observed over 18 preselected position angles. We present our method for reducing this data set into a CME catalogue. The resulting catalogue consists of 144 CMEs over the period January 2007 to February 2010, of which 110 were observed by STEREO-A and 77 were observed by STEREO-B. For each CME, the time-elongation profiles generated by the citizen scientists are averaged into a consensus profile along each position angle that the event was tracked. We consider this catalogue to be unique, being at present the only citizen science-generated CME catalogue, tracking CMEs over an elongation range of 4° out to a maximum of approximately 70°. Using single spacecraft fitting techniques, we estimate the speed, direction, solar source region, and latitudinal width of each CME. This shows that at present, the Solar Stormwatch catalogue (which covers only solar minimum years) contains almost exclusively slow CMEs, with a mean speed of approximately 350 km s^{-1}. The full catalogue is available for public access at www.met.reading.ac.uk/~spate/solarstormwatch. This includes, for each event, the unprocessed time-elongation profiles generated by Solar Stormwatch, the consensus time-elongation profiles, and a set of summary plots, as well as the estimated CME properties.

1. Introduction

Coronal mass ejections (CMEs) are eruptions of predominantly coronal plasma and magnetic flux out into the heliosphere [e.g., *Webb and Howard*, 2012] and are widely recognized as a key driver of space weather [*Hapgood*, 2011; *Cannon et al.*, 2013]. Earth impacting CMEs can be highly "geo-effective," potentially causing strong geomagnetic storms [*Gonzalez et al.*, 2001; *Borovsky and Denton*, 2006], and sufficiently energetic CMEs can also be a source of solar energetic particles, which pose a significant radiation hazard [*Barnard and Lockwood*, 2011; *Reames*, 2013]. *Hapgood* [2011] explains how, over approximately the last 150 years, society has grown increasingly dependent on technological systems that are vulnerable to the effects of disturbed space weather. A report from the Space Studies Board of the National Research Council in the United States suggests that severe space weather events would be very damaging for modern society [*National Research Council–Space Science Board*, 2008]. To put this in context, since 2011, "severe space weather" has been included in the United Kingdom's National Risk Register, where it is estimated that such events are as probable and disruptive as periods of "low temperature and heavy snow" and "heat waves" [*UK Cabinet Office*, 2013]. The effective mitigation of the risks associated with space weather hazards requires continued research into space weather-related phenomena, including developing a thorough understanding of the physics of CMEs.

Research into CMEs has been and will continue to be facilitated by an increasing set of CME observations, which can be split into two categories: remote sensing and in situ. A thorough description of the history of

©2014. The Authors.

CME observations is provided by *Webb and Howard* [2012], and here we focus on remote sensing observations. Modern remote sensing observations consist primarily of white-light coronagraph images, such as those taken by the Large Angle Spectroscopic Coronagraph (LASCO) on the SOHO satellite [*Brueckner et al.*, 1995], the COR Lyot coronagraphs on the twin STEREO satellites [*Howard et al.*, 2008], and also wide-angle Heliospheric Imagers, such as the Solar Mass Ejection Imager (SMEI) on the Coriolis satellite [*Webb et al.*, 2006] and the Heliospheric Imagers (HI) on the STEREO satellites [*Eyles et al.*, 2008]. Many of these observations are complimentary as they cover different regions of the solar corona and inner heliosphere and also span different periods of time.

It is common for these data sets to be reduced into more readily useable catalogues of CMEs, providing estimates of key CME properties such as the propagation direction, speed, angular width, and mass. Any catalogue is constructed by applying a CME identification algorithm to a set of observations; consequently, there is variability between the catalogues which depends on the instrumentation and data sets employed and the specific algorithm used to identify CMEs. An important difference between CME identification algorithms is whether or not they are applied manually, by a scientist analyzing the data, or automatically, as a set of codified rules which can be applied by a computer.

Some well-known examples (but by no means an exhaustive list) of such catalogues are Coordinated Data Analysis Workshop (CDAW) [*Gopalswamy et al.*, 2009], the SMEI list [*Webb et al.*, 2006], CACTus [*Robbrecht et al.*, 2009], CORIMP [*Byrne et al.*, 2012; *Morgan et al.*, 2012], AICMED [*Tappin et al.*, 2012], ARTEMIS [*Floyd et al.*, 2013], and SEEDS [*Olmedo et al.*, 2008]. The CDAW and SMEI catalogues both manually identify CMEs in the LASCO and SMEI data sets, respectively. CACTus, CORIMP, ARTEMIS, and SEEDS all employ different automated algorithms to identify CMEs in LASCO images, and the CACTus and CORIMP methodologies are also used with the COR data from the STEREO satellites. AICMED is an automated routine applied to the SMEI data, although the authors of that work note that in its present form manual inspection is also required to ensure the reliability of the results [*Tappin et al.*, 2012].

Yashiro et al. [2008] discussed the strengths and challenges of manual and automated methods of CME detection by comparing the CDAW and CACTus catalogues. A fundamental problem with manual identification of CMEs is that the definition of a CME is subjective and variable in time due to the limitations of individual judgment. Also, a more practical problem is that manual identification can be a heavy time burden on an individual or small team. Automated routines can, to some extent, avoid these problems. For example, given the same input data, they should yield repeatable results as, although the CME definition is still subjective, it is absolute. Such automated routines are also much more time efficient (at least after the development stage). However, as an example of the type of problem that automated routines can encounter, the specific comparison between CDAW and CACTus performed by *Yashiro et al.* [2008] revealed that CACTus was misidentifying many fast CMEs, presumably due to the difficulty of creating definitive rules to define such a variable phenomenon as a CME.

In this paper we report on a novel means of building a CME catalogue which uses a method that is a middle ground between the manual and automated systems discussed previously. The catalogue is derived from data from the Heliospheric Imagers aboard both STEREO satellites, and the CMEs are identified and tracked through the heliosphere as part of a citizen science project, Solar Stormwatch (SSW) (http://www.solarstormwatch.com), which was the first citizen science project to concentrate on space weather science and is a member of the Zooniverse set of citizen science projects [*Fortson et al.*, 2011]. Solar Stormwatch consists of several activities, available via a web interface, in which volunteers (presently >16,000) both identify and track CMEs through the HI fields of view (FOV). These results can then be statistically reduced into a consolidated set of CME observations. This method has several advantages over manual detection by a single observer, as, with a large number of manual identifications of single events, we are able to derive an average profile for each event and also estimate an uncertainty of this average profile, which helps mitigate the subjective biases of manual identification and is much less of a time burden on individuals.

In section 2, we briefly review the data used by Solar Stormwatch, namely, the HI observations from the STEREO spacecraft. In section 3, we summarize the design and operation of the Solar Stormwatch project. Section 4 details the data processing required to reduce the citizen scientists' (CSs) observations down to the resulting CME catalogue. In section 5, we review the CME catalogue. Finally, in section 6, we will discuss some summary statistics of the estimated CME properties.

©2014. The Authors.

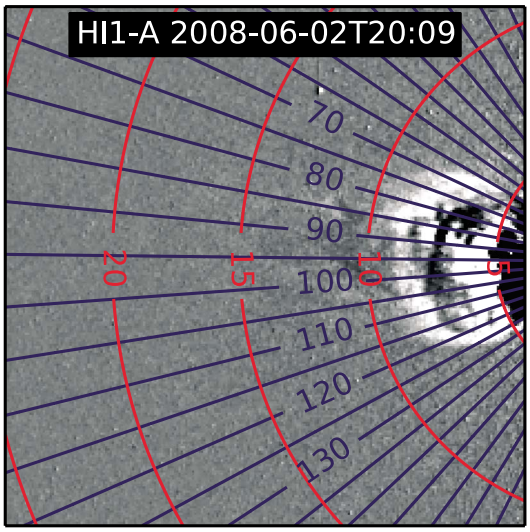

Figure 1. An example of a differenced image from the HI1A camera, overlaid with contours of constant position angles (PA) (in blue) and constant elongation angle (in red). The elongation and PA contours are in 5° increments. A CME is visible to the right of the image, between 5° and 10° elongation, and with maximum extent in PA between 65° and 135°.

2. Review of the STEREO Mission

The twin Solar Terrestrial Relations Observatory (STEREO) spacecraft were launched in 2006 into Earth-like heliocentric orbits, one ahead (STEREO A: STA) and one behind (STEREO B: STB) the Earth. The orbital speeds of the two spacecraft result in their gradual separation, with each drifting away from the Earth at a rate that increases the spacecraft-Sun-Earth angle by approximately 22.5° per year. Each STEREO spacecraft carries the Sun-Earth Connection Coronal Heliospheric Investigation suite of imaging instrumentation including an Extreme Ultraviolet Imager (EUVI), two white-light coronagraphs (COR1 and COR2), and a Heliospheric Imager [*Howard et al.*, 2008]. Each HI instrument contains two wide-field white-light cameras (HI1 and HI2) that can image solar wind structures such as CMEs and was the first to observe the signatures of corotating interaction regions (CIRs) over an elongation angle range from 4° to 88° from the Sun [*Rouillard et al.*, 2008; *Sheeley et al.*, 2008]. HI1 has a 20° FOV, extending from 4° to 24° in the ecliptic plane with a nominal image cadence of 40 min while HI2 has a 70° FOV from 18.8° to 88.7° in the ecliptic plane, with a nominal image cadence of 120 min [*Eyles et al.*, 2008]. The FOV of both HI1 and HI2 are nominally centered in the ecliptic plane.

In the HI images, the solar wind and the density structures within it are observed via sunlight that has undergone Thomson scattering from free electrons in the solar wind plasma. However, the majority of the signal received by the cameras results from light scattered from interplanetary dust (the *F* corona) and this needs to be subtracted from the images before the solar wind transient features can be seen. Since the *F* corona is slowly varying and does not move significantly relative to the HI FOV for a given epoch, it can be characterized over a small sequence of images and subtracted from each image within this epoch to generate background-subtracted images that reveal features within the solar wind. Alternatively, consecutive images can be subtracted from each other. In this way, relatively static features, such as the *F* corona, are removed, while transient enhancements and depletions in the electron density appear as brighter and darker features, respectively. Such "differenced" images have the advantage that they require little data processing but also have the disadvantage that they have a rather abstract appearance with any transient features being associated with both light and dark patches resulting from their movement between frames (an example of a differenced image can be seen in Figure 1). From their unique position outside the Sun-Earth line, the STEREO spacecraft can and have been used to track solar wind structures from the Sun's atmosphere out to 1 AU and beyond, including those directed toward Earth [*Davis et al.*, 2009].

The highest resolution images from the spacecraft are down linked once per day via the deep-space network [*Howard et al.*, 2008]. Such data are processed on the ground and made available several days after the images were taken. The STEREO spacecraft also broadcast a continuous stream of low-resolution data via a space weather beacon that is gathered by a network of ground stations on a best effort basis. While this latter data stream is less complete and of lower resolution, it is made available within a few hours of being collected and as such enables near-real-time analysis of the data. SSW makes use of both the higher-resolution science data and the space weather beacon mode data.

The wide field imaging capabilities of HI mean that it is most convenient to discuss the location of features in the cameras' FOVs in terms of elongation angles (ϵ) and position angles (PAs). The elongation angle of a target is equal to the angle between the observer-Sun center vector and observer-target vector, and the PA is equal to the angle in the image plane between the target-Sun center vector and the direction of

©2014. The Authors.

Solar-North, in an anticlockwise sense. Figure 1 demonstrates this with an example of a differenced image from HI1A, over which contours of constant PA (in blue) and elongation (in red) have been plotted.

Techniques have been developed that allow us to estimate the speed and trajectory of solar wind transients in HI images [*Sheeley et al.*, 1999, 2008; *Rouillard et al.*, 2008], and these have been used to track Earth-impacting CMEs both retrospectively [*Davis et al.*, 2009] and in near real time [*Davis et al.*, 2011]. This requires identifying the transient's time-elongation (t-ϵ) profile along a constant PA. Often, the t-ϵ profile is extracted from a t-ϵ map, colloquially known as a J-map. J-maps are constructed from a sequence of images by extracting the brightness profile as a function of elongation, averaged over a narrow range of PAs, and stacking these vertically as a function of time (on the x axis). In such J-maps, antisunward propagating transients have positive gradients and an example of a J-map, built from both HI1 and HI2 images, can be seen in Figure 3. *Davies et al.* [2009] provide a full account of the construction of J-maps.

Extracting t-ϵ profiles from J-maps is labor intensive and time consuming, resulting in research that targets individual events [*Rouillard et al.*, 2009] or a limited survey of such events [*Davis et al.*, 2010; *Möstl et al.*, 2014]. With the HI instruments alone gathering over 35,000 images per year, the STEREO data set contains far more information about CMEs and other solar wind transients than can be easily analyzed in detail by the limited number of specialist researchers in the field.

SSW makes use of J-maps constructed from both HI1 and HI2 differenced images along multiple PAs. Specifically, J-maps were created along 18 different PAs, in 5° increments, spanning 50°–130° for STA and 230°–310° for STB, with the final J-maps in each 18-member set being created from a 5° band centered on the ecliptic plane. In each J-map the HI1 and HI2 elongation profiles are joined at $\epsilon = 18.8°$.

The SSW results have been generated from HI data spanning the period from January 2007 to February 2010. Over the duration of this period, the separation between STA and STB has increased from less than 1° of longitude (in Heliocentric Earth Ecliptic coordinates) to 137°, while the separation of STA and STB from Earth increased from less than 1° to 65° and 71°, respectively.

3. The Solar Stormwatch Project

3.1. Design and Operation

Solar Stormwatch was conceived and built by a collaboration between the Rutherford Appleton Laboratory, the Royal Observatory Greenwich (ROG, part of Royal Museums Greenwich), and Zooniverse at the University of Oxford. To be consistent with the principles of the Citizen Science Alliance (http://www.citizensciencealliance.org), SSW was designed from the outset to produce academic research and also a rewarding experience for the participants. The SSW website was designed by the digital media team at the ROG. In particular, the ROG was responsible for designing the interface to each data analysis task, the website's overall styling and producing its multimedia content. To help develop the user experience, the game designers Six to Start (http://sixtostart.com) were consulted on the ways that game mechanics could be used to motivate volunteers, as analysis of prior citizen science and crowd sourcing projects has revealed that this can be beneficial [*Holley*, 2009; *Raddick et al.*, 2010]. The final format of the SSW interface was the result of an iterative design procedure, where initial ideas and interfaces were tested on potential users, in this case visitors to the ROG. This process continued until the user interface and each activity was deemed easy to use for volunteers and produced the results the science team needed. Each activity has an associated introductory training exercise to help ensure that the participants are able to accurately perform the required tasks. Additional details about the design and operation of the Zooniverse platform, which was used to construct Solar Stormwatch, are provided in section A1.

3.2. Solar Stormwatch Activities

Below we provide a summary of the two SSW activities used in this study, *Spot!* and *Trace-it!*. However, there are four other SSW activities: the data from which has been used to investigate interplanetary dust distributions [*Davis et al.*, 2012a], the geometry of CMEs [*Savani et al.*, 2012], the prediction of arrival times of high speed solar wind streams at Earth [*Davis et al.*, 2012b], and also in the validation of CME arrival predictions using the real-time HI observations (K. Tucker-Hood et al., submitted manuscript, 2014). These other activities are summarized in the section A2.

3.2.1. Spot!

The simplest SSW activity is the identification of CMEs within the HI1 images and the making of an initial estimate of the CME trajectory. Movies, created from STA and STB HI1 images, are presented simultaneously;

©2014. The Authors.

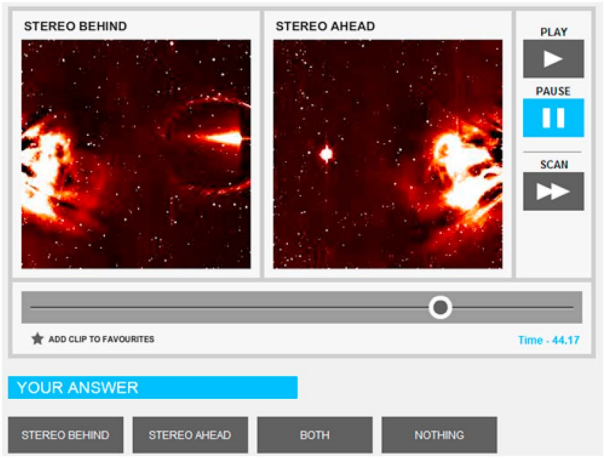

Figure 2. An image of the user interface for the Spot! activity. The left and right hand images show simultaneous frames from the HI1B and HI1A movies, respectively. In this instance, the movie has been paused when a CME is visible in the FOVs of both HI1 cameras. Users are required to identify the occurrence of CMEs in either or both of the movies. If a CME is present, users are then required to estimate the times at which the CME first enters, and is halfway across, the HI1 FOV.

each movie included 14 days of data, but successive movies overlapped by 7 days to reduce the chance of CMEs near the end of movies being missed. Figure 2 shows an example of the user interface for this activity. Participants were asked to view the movies and identify CMEs seen from either one or both spacecraft. Those CMEs that appeared in both cameras indicated an event that was heading between the spacecraft and therefore roughly toward Earth. For such events, participants were asked to record the time at which the CME front reached a fixed point in the HI-B FOV. The simultaneous elongation of the CME front (along the ecliptic plane) in the HI-A FOV was then recorded by drawing a sliding bar across to mark the CME front. Participants were then asked to step backward through the movies and note the times at which the CME entered each spacecraft's FOV. From these measurements and a knowledge of the spacecraft position at the time of the observation, it was possible to use these stereoscopic observations of the CME front to make initial estimates of speeds and trajectories (assuming a constant speed and direction) for such CMEs.

For those CMEs seen by only one camera, it was likely that the trajectory of the storm was such that it did not travel between the spacecraft or was directed much more toward one craft than the other. In such cases, participants were invited to log the frame in which the CME had reached the marked location and also the frame in which the CME entered the FOV. While no true speed and trajectory can be calculated through triangulation for such events, their start times were noted in a similar manner to the start times of events seen in both cameras.

To identify probable CMEs, the set of estimated CME start times generated by the CSs were searched for clusters. A cluster in the estimated start times means that many CSs all observed a CME entering the HI FOV at a similar time. Clusters are calculated by computing the number of estimates of CME start times as a function of time, using a 12 h sliding window, stepped by 2 h, and applying a threshold of 20 counts. Local maxima above this threshold define periods when CMEs are first observable within the HI FOV. This information was used to guide participants in the following Trace-it! task by directing them to segments in J-maps where it is likely a CME can be observed.

3.2.2. Trace-It!

The Trace-it! activity was designed to enable a more detailed analysis of the CMEs identified in the initial Spot! activity and requires participants to manually track the propagation of CMEs through the set of combined HI1/HI2 difference J-maps described at the end of section 2. Figure 3 shows an example of the user interface for this activity. The horizontal blue bar at the bottom of the J-map marks the region in which the Spot! results suggest that a CME first appeared in the HI1 FOV. Participants are required to place up to 20 markers, charting the trajectory of the CME through the J-map. The accuracy with which the speed and trajectory of a CME can be determined from a single t-ϵ profile (see section 6 for a description of methods to calculate this) increases with the maximum elongation extent of the profile [*Williams et al.*, 2009], and so participants are directed to track CMEs out beyond 35° wherever possible, which is marked by the horizontal blue dashed line. When finished, each t-ϵ profile is saved to a database which forms the bulk of the data used in this work.

The motivation behind directing participants to regions of J-maps identified from the Spot! results was to minimize the number of users profiling the trajectories of other solar transients, such as the isolated plasma parcels that can be seen moving at the stream interface associated with CIRs [*Rouillard et al.*, 2008, 2010]. However, as will be discussed later, it has been found that although participants were frequently sensibly

©2014. The Authors.

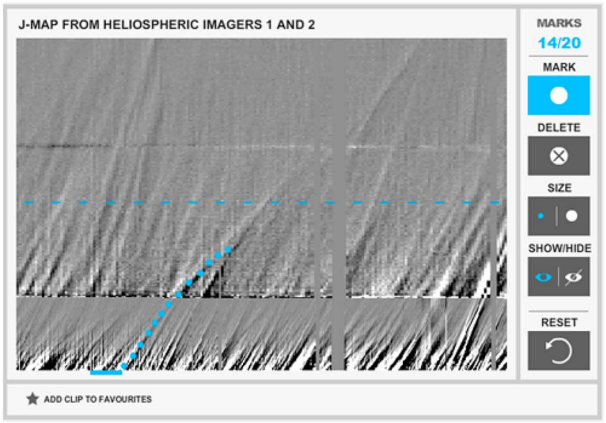

Figure 3. An image of the user interface for the Trace-it! activity. The central figure is a J-map formed from HI1 data (bottom portion) and HI2 data (top portion), as described in section 2. The (horizontal) time axis spans a period of approximately 14 days, and the (vertical) elongation axis spans a range of approximately 70°. The horizontal blue bar at the bottom of the J-map marks a range of times in which a CME has been observed to enter the HI1 FOV in the Spot! activity, while the blue dots mark the user-selected t-ϵ profile for a transient visible in this J-map. The horizontal blue dashed line marks an elongation of 35°; users are encouraged to try to track the CME profiles out past this elongation.

tracking the profiles of solar wind transients, quite often these were not the profiles that they had been directed to by Spot! activity.

4. Data Processing

In the following section, we detail the processing used to reduce the large database of t-ϵ profiles generated by the Trace-it! activity down into the resulting CME catalogue. This process has several stages; the initial quality control of the t-ϵ profiles, the grouping of the t-ϵ profiles into individual events, further quality control to correct spurious associations, and, for each event, the averaging of the individual t-ϵ profiles along each PA into consensus profiles.

4.1. Initial Quality Control

Every t-ϵ profile generated by a citizen scientist (hereafter CS) is stored in a database, and each entry includes the following information: the spacecraft ID, the PA of the J-map that a CS was analyzing, the CS user ID number, and the t-ϵ profile, the latter of which is stored as a set of decimal Julian dates and elongation angle in degrees. The first stage of processing these data involves applying a set of basic quality control rules to identify and remove erroneous profiles. Such errors seem to have several causes. Amongst the most common are caused by the presence of large discontinuities in profiles due to the CS tracking more than one feature in a J-map but incorrectly submitting the combined profiles as one event. Another common source of error is due to CS tracking the orbit of a planet, rather than a CME, through the FOV of the J-map. We draw attention to these errors in particular as it seems reasonable to suggest that both could be reduced in frequency or may be avoided completely by fairly minor modifications to the Trace-it! user interface. For example, planet and comet trajectories are predictable and so could potentially be masked out of the J-maps.

At this stage we apply eight rules to the data which are that a t-ϵ profile must not (1) have less than three points, (2) span more than 10 days, (3) span less than 5° in elongation, (4) have more than a 5 day discontinuity in the sorted time sequence, (5) have more than a 10° discontinuity in the elongation sequence sorted in time, (6) must not start above 20° elongation, (7) must not be contained below 10°, and (8) the times and elongations must be positively correlated (a negative correlation would imply sunward propagation). These rules were derived through the inspection of a subset of the profiles and comparison with the associated J-maps.

At the time of processing, there were 38,171 profiles in total, and this level of quality control removes 5988, leaving 32,183 profiles for further analysis, 18,829 from STA and 13,354 from STB. Table 1 provides a breakdown of the number of events rejected for each reason. Note that the sum of the rejections by reason does not equal the total number rejected, as frequently t-ϵ profiles fail for multiple reasons. The initial 38,171 profiles were generated by 4599 unique users, whereas the remaining 32,183 useable profiles were generated by 2634. Therefore, ≈ 50% of the participants that

Table 1. Breakdown of Rejected Profiles By Reason

Reason	# of Tracks
< 3 points	481
Range of $t > 10$ days	2261
Range of $\epsilon < 5°$	1776
Discontinuity in t	700
Discontinuity in e	2936
All $\epsilon > 20°$	3269
All $\epsilon < 10°$	279
Corr$(t, \epsilon) < 0$	1899

©2014. The Authors.

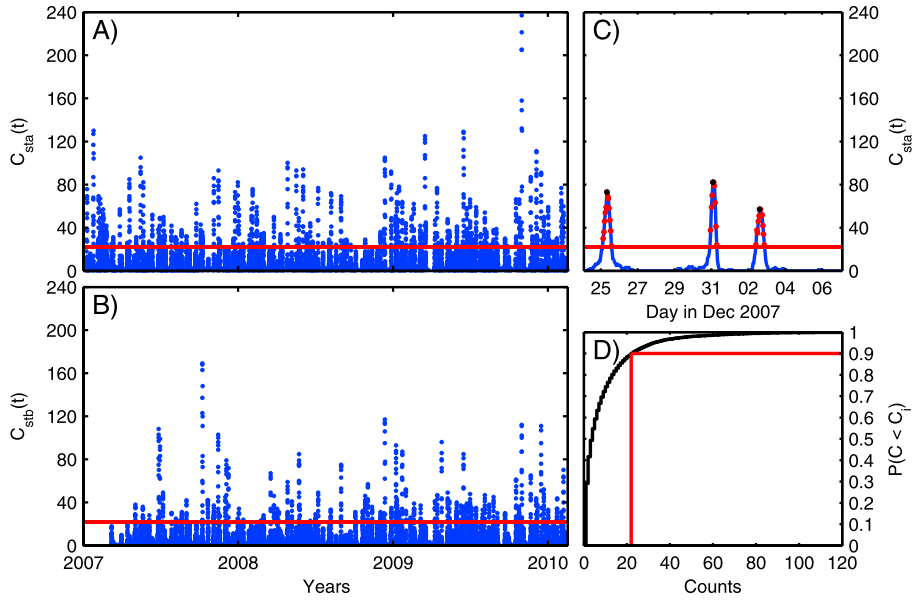

Figure 4. (a) The hourly count rate of CS-identified t-ϵ profiles that begin in each UT hour, for all PAs, in STA, C_{sta} (blue dots). The red line marks the threshold used to identify CMEs. (b) The same as Figure 4a but for C_{stb} from STB. (c) An example of a short period of the C_{sta} variation (blue line) which shows the maxima defining the onset of three CMEs, with the red dots showing the points above the event threshold and the maximum shown by the black dots. (d) The empirical cumulative distribution of the pooled, nonzero, C_{sta} and C_{stb} counts, used to define the event threshold, marked by the red lines in each panel (black line).

engaged with the SSW project generated >80% of the profiles that meet the requirement of obeying all of our eight rules.

Although it is possible that some of the rejected profiles could be salvaged, in particular by identifying those which relate to more than one feature and separating them, with the number of profiles remaining for analysis, it was considered an unnecessary complication to build this into the data processing at this stage but is a possible future improvement.

4.2. Profile Clustering

4.2.1. A Fixed Window Around Spot! CME Times

We initially tried to group the t-ϵ profiles into individual CMEs by comparison with the CME times identified in Spot!. Over the period analyzed, the use of Spot! resulted in the identification of 145 CMEs in STA and 113 in STB, 50 of which were seen by both spacecraft. To associate the t-ϵ profiles with the Spot! CME times, we searched for every t-ϵ profile that began within a fixed window of the Spot! CME start time. However, there were complications with this approach which meant that it would have been an inefficient use of the Trace-it! data. This leads us to take a different approach, which is discussed below. The details of our investigation into this method are not central to the results of this work and are quite long. We therefore include these details in Appendix B.

4.2.2. Clustering the Trace-It! t-ϵ Profiles

To identify groups of t-ϵ profiles associated with individual CMEs, we search for clusters in the start times and PAs of the profiles for STA and STB separately. This is achieved by calculating the total number of profiles that begin in a running 7 h window for each UT hour spanned by the data sets from 01-01-2007 to 17-02-2010 for STA and from 08-03-2007 to 17-02-2010 for STB. This running count of start times is referred to as $C_{sta}(t)$ and $C_{stb}(t)$ for STA and STB, respectively, and is plotted as blue dots in Figures 4a and 4b. Maxima in the $C_{sta}(t)$ and $C_{stb}(t)$ time series occur when many users observe features in the J-maps at multiple PAs at similar times. Here the maxima are defined as all periods where the counts are greater than a constant threshold for more than three consecutive hours. The threshold was chosen to be the 90th percentile of the empirical distribution of the nonzero, pooled $C_{sta}(t)$ and $C_{stb}(t)$ counts, which is equal to 22 counts; this distribution and the selected threshold can be seen in Figure 4d. Figure 4c shows an example of a short interval of $C_{sta}(t)$ for a 12 day period starting on 24-12-2007, which shows in detail three of the maxima in $C_{sta}(t)$, where the red dots show the hourly values above the threshold and the black dots mark the peaks of

©2014. The Authors.

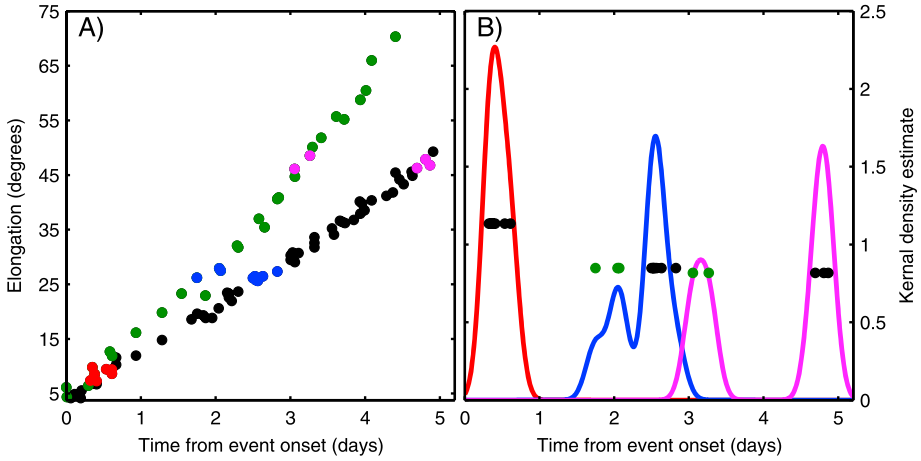

Figure 5. An example to illustrate the method used to identify outlying t-ϵ profiles for a CME tracked in STA along a PA of 70°. (a) The primary cluster of t-ϵ profiles (black dots) and the t-ϵ profiles identified as outliers (green dots). The red, blue, and magenta groupings identify the points in three nonconsecutive example elongation bins. (b) The red, blue, and magenta lines give the kernel density estimates of the density of points in the corresponding elongation bins from Figure 5a, while the black and green dots show the locations of points identified as belonging to the primary cluster and as outliers, respectively.

the maxima. Defining each of the maxima in $C_{sta}(t)$ and $C_{stb}(t)$ as an event, we associate the corresponding set of t-ϵ profiles with that event (i.e., those that start within ±3 h of the black dots in Figure 4c).

We then analyze the distribution of t-ϵ profiles as a function of PA over each event. To have confidence that features have been tracked robustly, we disregard any PA which has less than three t-ϵ profiles associated with it. Furthermore, we disregard any PA which is separated by more than 20° from the main cluster of PAs so that profiles at widely separated PAs are not incorrectly associated with the events. After these PAs have been discarded, the total number of t-ϵ profiles contributing to the event is recalculated and the event is discarded if the number falls below the event threshold of 22 profiles. There is no constraint on the minimum number of PAs an event must be observed over.

Using the discussed criteria defines 115 clusters in STA and 79 clusters in STB. In total, these clusters include 10,368 t-ϵ profiles, 6301 from STA, and 4067 from STB so that ≈ 32% of the profiles have been attributed to events. The smaller number of events observed by STB is likely due to the fact that the HI images from STB are frequently noisier, due to, for example, dust impacts affecting HI-B images more than HI-A images, and so it is often harder to identify transients in the HI-B J-maps [*Davis et al.*, 2012a].

4.3. Clustering Quality Control

Initial analysis of the clusters of profiles identified as individual events in the previous section showed that a further level of quality control was required before sensible consensus profiles could be constructed for each event. The reason for this is that there are t-ϵ profiles that diverge from the primary cluster of profiles. An example of this is shown in Figure 5a, where the black dots show the primary cluster of profiles and the green dots show two profiles that have been identified as outliers by a method explained below. Such instances usually occur because identifying the t-ϵ of a CME within a particularly complex coronal outflow can sometimes be ambiguous.

We employ the following algorithm to identify the outlying profiles. Considering each observed PA of each event, in turn, we

1. Bin all points in the set of profiles into 3° wide elongation bins, spanning the whole elongation range. Examples of this, for three elongation bins, are shown by the red, blue and magenta groups in Figure 5a.
2. For each elongation bin, the density of points as a function of time is estimated, using a kernel density estimator. This works by centering a normal distribution at the time coordinate of each point, with a standard deviation of 3 h, and then the density of points as a function of time is estimated as the sum of

©2014. The Authors.

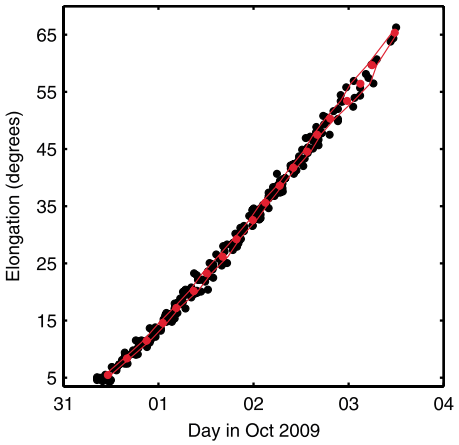

Figure 6. An example of a consensus t-ϵ profile for CME number 99 from STA tracked along a PA of 115°. The black dots show the individual t-ϵ profiles, and the red dots mark the consensus profile while the two red lines indicate the uncertainty in the mean time coordinates, defined here as being equal to two standard errors in the mean time.

the individual normal distributions. Examples of this are shown by the red, blue, and magenta curves in Figure 5b.

3. The primary cluster in each elongation bin is identified by the global maximum of the density profile in that elongation bin.
4. The full width at half maximum (FWHM) of the primary cluster is calculated, and any points that fall outside of the time of the global maximum ± FWHM are marked as points belonging to potentially outlying profiles.
5. Any profile that is identified as including points that lie outside the primary cluster in three or more elongation bins is regarded as an outlier and removed from the event.
6. The event is checked to ensure it still meets the criteria used to define events in section 4.2.2 and is discarded if it does not.

A further 510 profiles were identified as outliers and removed, 325 from STA and 185 from STB. Five events from STA and two events from STB were then discarded as they fell below the threshold of 22 profiles. As a result, in total, 5858 profiles have been associated to the 110 events seen by STA and 3832 profiles have been associated to the 77 events seen by STB. A detailed description of the composition of the final SSW catalogue is provided in section 5. So that the data processing employed here is transparent, a plot has been generated each time a profile has been identified as an outlier and these are available for viewing as part of the online database.

4.4. Consensus Profiles

At this point, each event has a set of at least three t-ϵ profiles along each PA on which the event was observed. These are now averaged to provide a consensus profile and an uncertainty estimate along each PA. The consensus profile is calculated using the same elongation bins as are used in section 4.3 (3° wide elongation bins, spanning the whole elongation range). In each elongation bin, the mean of the time coordinates and the mean of the elongation coordinates are calculated, as well as the standard error on the mean (SEM) for each quantity; the uncertainties in the mean time and elongation coordinates are calculated as $2 \times$ SEM. Figure 6 shows an example of a consensus t-ϵ profile, for CME number 90 from STA, tracked along a PA of 115°. The black dots show the individual t-ϵ profiles, and the red dots mark the consensus profile; the two red lines mark the uncertainty in the mean time coordinates.

4.5. Imaging the CME Fronts

So far, we have considered the evolution of an event as represented by a cluster of t-ϵ profiles over multiple PAs, but it has not yet been demonstrated that this methodology can accurately capture the shape of the CME fronts that have been visually identified in the HI images. Figure 7 displays an illustrative test of how successfully CMEs can be tracked in the SSW catalogue. This set of images shows the propagation of two different CMEs through the HI1A FOV. Each event is shown in both the science images and also differenced science images, like those used to form the J-maps in which they were tracked; the first event (event number 49) occupies columns A and B, and the second event (event number 99) occupies columns C and D. Time increases downward, and each frame is separated by approximately 6 h. The yellow lines mark the maximum and minimum PAs used in constructing the J-maps served in Trace-it!, outside of which nothing can be tracked. The location of the CME fronts are overlaid on each image as regions bounded by red lines. These are calculated by interpolating the consensus t-ϵ profiles (see Figure 6) at each PA that the CME was tracked. The elongation-width of the bounded region represents the uncertainty in the consensus profile at that PA and time. In each instance, the leading edge of CME has been independently identified as it propagates through the HI1A FOV. In most instances, the identified CME leading edge lies inside the red bounded region derived from the SSW results. Comparing events 49 and 99, it is clear that leading edge of event 99 has a more complicated structure. In particular, we draw attention to

©2014. The Authors.

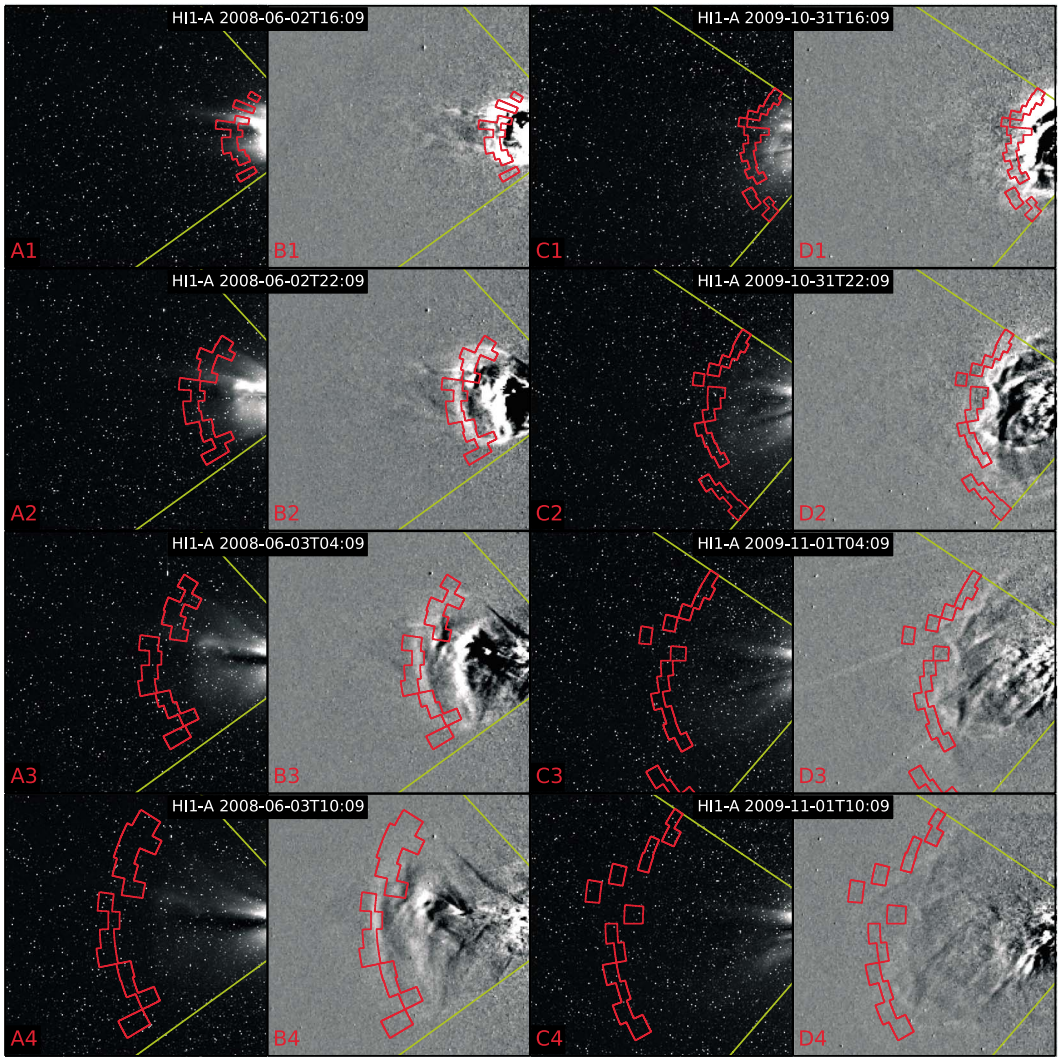

Figure 7. This set of images shows the propagation of two different SSW-identified CMEs (STA events 49 and 99) through the HI1A fields of view. Each event is shown in both the background-subtracted images (49: A1–A4, 99: C1-C4) and also differenced images (49: B1–B4, 99: D1–D4) (the latter being the source of the J-maps in which they were tracked). Time increases downward with each frame separated by approximately 6 h. The yellow lines mark the limits of the PAs used in the J-maps analyzed in Trace-it!. The locations of the CME fronts are overlaid on each image as regions bounded by red lines. These are calculated by interpolating the consensus t-ϵ profiles (see Figure 6) at each PA the CME was tracked. The elongation width of the bounded region represents the uncertainty in the consensus profiles.

the small-scale depression in the elongation extent of the central portion of the leading edge, spanning approximately 15° in PA; even this small-scale detail is resolvable within the errors of the consensus profiles.

There are of course limitations to the SSW catalogue. To highlight this, we consider STA events 63 and 64, which entered the HI1A FOV approximately 21 h apart. The propagation of these events through the HI1A FOV can be viewed in the movie provided as supporting information to this work. This movie has a similar format to Figures 7b and 7d, showing a sequence of differenced images from HI1A. Overlaid on these are bounded regions marking the CME fronts identified from the SSW results; event 63 is shown in red, and event 64 is shown in blue. The bulk of event 63 is tracked reasonably well, although the top portion of the CME front (at low PAs) is seemingly missed, whereas the opposite is true for event 64, the front of which is poorly tracked except at low PAs. We surmise that this is due to the close proximity of the t-ϵ profiles corresponding to each event in the J-maps. Event 63 is more clearly visible over most PAs, and it is only at the lower PAs where event 64 leaves an easily identifiable trace. This situation is further complicated by a short data gap (three missing images) near the onset of event 64, which makes it more difficult to obtain a

©2014. The Authors.

Table 2. Summary of Association of t-ϵ Profiles and Users With CMEs

	STA	STB
	For Each CME	
Min # of profiles	22	22
Mean # of profiles	53	50
Max # of profiles	243	198
Min # of CS	9	10
Mean # of CS	27	26
Max # of CS	100	93
	For Each Position Angle	
Min # of profiles	3	3
Mean # of profiles	7	7
Max # of profiles	22	22
Min # of CS	1	1
Mean # of CS	5	5
Max # of CS	17	17

sensible t-ϵ profile. Such short waiting times between events are quite rare in this catalogue, with a total of 12 events from the STA and STB lists having waiting times $\leq$ 21 h. The mean waiting time is 10.3 days for the STA list and 13.3 days for the STB list.

5. The Solar Stormwatch Catalogue

In this section, we provide a summary of the SSW CME catalogue. Both the raw and consensus t-ϵ profiles for each event, along each observed PA, are available to download at http://www.met.reading.ac.uk/~spate/stormwatch as formatted text files, as well as summary plots similar to Figure 6. In addition to this, plots generated as part of the clustering quality control (section 4.3) are available to browse so that it is clear what data have been selected and rejected.

5.1. Total Number of Events

In total, we have extracted 110 CMEs observed by STA and 77 CMEs observed by STB from the Trace-it! data. The total number of unique events in the SSW catalogue is fewer than 187, as some CMEs are common to both the STA and STB lists. We have estimated the total number of unique CMEs by searching for events which occur approximately coincidentally in time. The minimum waiting time between successive CME events is 10 h in the STA list and 14 h in the STB list. We looked for events in the STB list with onset times separated by less than 10 h from the onset time of an event in the STA list. It was found that 43 events are potentially common to both lists, such that the number of unique CMEs in the joint STA and STB catalogue is approximately 144 events. In the future, we plan to make a more robust estimate by also comparing the estimated source regions of the events.

5.2. Association With Spot! CME Times

We have also compared our STA and STB catalogues with the list of Spot! CME times, using the same method of searching for Spot! CME times that occur within 10 h of the Trace-it! derived CME onset times. For STA, 77 (70%) CMEs could be matched with Spot! events, while for STB, 57 (74%) CMEs could be matched with Spot! events.

5.3. The Association of t-ϵ Profiles and CS With CMEs

Table 2 summarizes the statistics of t-ϵ profiles and CSs associated with each CME, both in total and along each PA. Out of the 32,183 t-ϵ profiles analyzed, only 9690 (30.1%) have been associated to a CME and so there is still probably unused information about other CMEs in the Trace-it! data. Of course, more of the Trace-it! data could have been used by lowering the threshold used to select events in section 4.2.2, which would have created a larger list of CMEs. However, this would have come at the expense of having confidence that the resulting clusters of t-ϵ profiles robustly represented CMEs. Our preference is to be conservative in defining the conditions required to identify a CME from the Trace-it! data. Periodically updating the CME database, as more Trace-it! data are added, means that more events will be identified using the present threshold with time. In total, 1254 CS contributed to the current set of t-ϵ profiles used in the catalogue, which is $\approx$ 30% of all those that took part in the Trace-it! activity.

5.4. The Tracked Elongation Range

Table 3 summarizes the maximum elongation extent that the CMEs were tracked out too for both the raw and consensus t-ϵ profiles considering all PAs. In the training exercise for Trace-it!, users are directed to attempt to track transients out past 35° elongation (see Figure 3). This was motivated by previous research that concluded that transients needed to be tracked out over $\approx$ 30° elongation to obtain accurate estimates of the CME's speed and direction with the fixed-phi geometrical model [*Williams et al.*, 2009]. Therefore, given this instruction and the data selection rules used in section 4.1, it is unsurprising to find that the minimum elongation extent of all the raw t-ϵ profiles that contributed to the final CME catalogue was 28.4°.

©2014. The Authors.

Table 3. Summary of the Maximum Elongation Extent of CMEs That Were Tracked

	STA	STB
	Raw t-ϵ Profiles	
Min ϵ_{max}	28.4°	34.7°
Mean ϵ_{max}	58.6°	57.0°
Max ϵ_{max}	74.0°	73.8°
	Consensus t-ϵ Profiles	
Min ϵ_{max}	12.9°	27.2°
Mean ϵ_{max}	53.0°	51.7°
Max ϵ_{max}	72.8°	72.3°

The mean extent of the consensus profiles is comparable with, but slightly less than, the mean extent of the raw profiles, both of which are > 50°.

5.5. CME Occurrence Frequency

Another quantity of interest is the CME occurrence frequency over the duration of the SSW catalogue. This was calculated in a set of 30 day wide, nonoverlapping, windows that span the duration of the SSW catalogue for STA and STB separately. In Figure 8 the time series of the STA and STB CME occurrence frequencies are shown as the solid blue and solid red lines, respectively. Furthermore, for both STA and STB, the average CME occurrence frequency over the duration of the SSW catalogue was calculated. This is equal to 0.096 ± 0.009 d^{-1} for STA and 0.072 ± 0.008 d^{-1} for STB, shown by the dashed blue and dashed red lines in Figure 8. The STA CME occurrence frequency is typically slightly higher than for STB, although there are periods where they match closely. There are points where they match identically, and this is to be expected as many of the CMEs are common to both STA and STB (as discussed in section 5.1). After late 2008, both begin to show an increase in CME occurrence, likely signifying the rise in solar activity associated with the onset of solar cycle 24.

5.6. Comparison With the RAL-HI Event List

The RAL-HI Event List (www.stereo.rl.ac.uk/HIEventList.html) is another catalogue of solar transients observed by HI. This list was created by an expert observer visually identifying solar transients in J-maps formed from both HI1 and HI2 differenced images, centered on a 5° PA band around the ecliptic plane. This list does not explicitly differentiate between the transient features in the J-maps made by CMEs and CIRs. Therefore, this catalogue contains both CMEs and CIRs.

There are significant differences between the RAL-HI event list and the SSW catalogue, which complicates a direct comparison. For example, the RAL-HI events are only tracked in the 5° PA band centered on the ecliptic plane, while the SSW events can be tracked over 17 distinct PAs, spanning 80° in total. Furthermore, many of the SSW events are linked with a visual identification of a CME in the HI images (see section 5.2), whereas the RAL-HI catalogue only identifies events from their transient profile in J-maps built from differenced images, which increases the visibility of faint features. Therefore, these catalogues have very different event definitions.

The RAL-HI catalogue contains approximately 7 times more events than the SSW catalogue, with 792 STA events and 468 STB events over the same period spanned by the SSW catalogue. This equates to mean occurrence frequencies of 0.692 ± 0.025 d^{-1} and 0.409 ± 0.019 d^{-1} for STA and STB, respectively. However, the increased RAL-HI event frequency must be viewed in the context that this list includes events which are narrower in PA than SSW would typically identify and also includes CIRs. On the level of individual events, the SSW catalogue contains more information, as it tracks events over multiple PAs, which allows us to estimate the physical size of the CME. Furthermore, the SSW catalogue includes events propagating outside of the ecliptic plane, which RAL-HI catalogue does not.

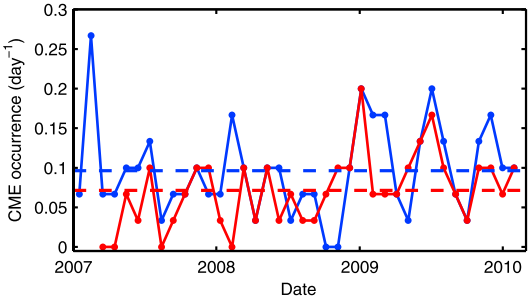

Figure 8. The time series of the CME occurrence frequency for STA (solid blue) and STB (solid red). These are calculated in nonoverlapping 30 day wide windows. Also shown are the average CME occurrence frequencies for STA and STB over the whole period marked by the blue dashed and red dashed lines, respectively.

Here we do not make a direct comparison with CME catalogues derived from coronagraph observations. However, we note that as the brightness of CMEs typically decreases with increasing heliocentric distance, CMEs are easier to observe in coronagraphs than in HI. Therefore, the SSW catalogue is likely to

©2014. The Authors.

contain only the biggest and brightest CMEs observed by coronagraphs and is unlikely to include any CMEs not seen by coronagraphs during normal operations. Furthermore, coronagraph-derived CME catalogues will also identify more CMEs due to observing a wider range of PAs than the HI instruments.

As far as we are aware, the SSW catalogue is the first CME catalogue which tracks CMEs over multiple position angles and over an elongation range spanning 4° to 70°. This allows us to study the structure and dynamics of CME fronts at elongation angles not accessible to coronagraphs and at position angles not considered by the RAL-HI list.

6. Estimating CME Properties

Several methods have been developed that allow estimation of the speed and trajectory of a CME from its t-ϵ profile observed from a single satellite. Three widely used methods are fixed-phi fitting (FPF) [*Sheeley et al.*, 1999, 2008; *Rouillard et al.*, 2008], harmonic mean fitting (HMF) [*Lugaz*, 2010], and self-similar expansion fitting (SSEF) [*Davies et al.*, 2012; *Möstl and Davies*, 2012]. These methods work by assuming a fixed geometry for the CME structure and by assuming that the CME propagates at a constant speed in a radial direction. For each model, a theoretical expression for the elongation angle as a function of time has been derived such that a numerical fit between the observed t-ϵ profile and the theoretical elongation angle variation can yield estimates of the CME speed (V_{cme}) and direction (ϕ). The FPF technique is the simplest of these three, modeling the CME as a point-source moving radially outward from the Sun at constant speed. The HMF technique models the CME as a radially expanding circle anchored at the Sun center. Finally, the SSEF technique models the CME as a radially expanding circle that is not anchored to the Sun, but subtends a fixed angle with respect to the Sun center. As such, the SSEF technique can model a continuum of CME geometries, for which the FPF and HMF techniques are two limiting cases. *Möstl et al.* [2014] recently reviewed the performance of these single-spacecraft-fitting methods and demonstrated that the FPF method provides the least biased estimate of the CME trajectory, with both the HMF and SSEF (using a CME half width of 45°) methods tending to give biased estimates of ϕ. Therefore, we have chosen to use the FPF method to estimate V_{cme} and ϕ for each CME in the catalogue. This is calculated using a least squares fit between the consensus t-ϵ profile corresponding to the central PA of the event and the theoretical equation for this geometry. We note that the minimum elongation extent for the consensus t-ϵ profiles of the central PAs is 34.9° for STA events and 33.1° for STB events. The estimated CME properties values are provided, also with error estimates calculated in the same way as done by *Rouillard et al.* [2010] and *Williams et al.* [2009], in the summary database of the catalogue. However, a point to note regarding the error estimates, as also discussed by many authors including *Savani et al.* [2012] and *Möstl et al.* [2014], is that these errors relate only to the quality of the fit between the observed and theoretical profiles. There is, for the FPF, HMF, and SSEF methods, an additional unquantified error which depends on how well the assumptions of each method apply to the event. We also estimate the solar source region and latitudinal width of the CME, both in Heliocentric Earth Equatorial coordinates, following the method of *Savani et al.* [2012].

6.1. The Distribution of CME Speeds and Latitudinal Widths

Figure 9 presents histograms of the distributions of the estimated CME speeds (Figure 9a) and latitudinal widths (Figure 9b). The distributions for STA and STB are shown separately as red and blue histograms, respectively, while the means of each distribution are shown by vertical dashed lines of the same color. These histograms are calculated using speed bins of 50 km s^{-1} and the latitudinal width bins of 7.5°. To allow a clearer scaling of the x axis, one event from the STA speed distribution lies outside the x axis limit, with an estimated speed of 1330 km s^{-1}. The mean CME speed is 365 ± 12 km s^{-1} for STA and 337 ± 6 km s^{-1} for STB. These speed distributions show that the SSW catalogue presently contains almost exclusively slow CMEs, with only one event from STA having an estimated speed >1000 km s^{-1}. This is consistent with these CMEs originating during the deep minimum in solar activity between the solar cycles 23 and 24. A comparison with the LASCO CDAW catalogue over the same period as spanned by the SSW catalogue revealed that LASCO contains 3157 CMEs, only 4 of which had estimated plane-of-sky speeds >1000 km s^{-1}. We note that LASCO images CMEs are lower in the solar atmosphere than HI. Therefore, it is probable that a fast CME observed by LASCO will have a lower speed when observed by HI, having been decelerated as it propagated outward by interaction with the solar wind.

Due to limitations in the HI FOV, the SSW system only uses J-maps spanning a limited range of 80° in PA; this will potentially limit estimates of the latitudinal extent of some CMEs. However, only six events from STA and

©2014. The Authors.

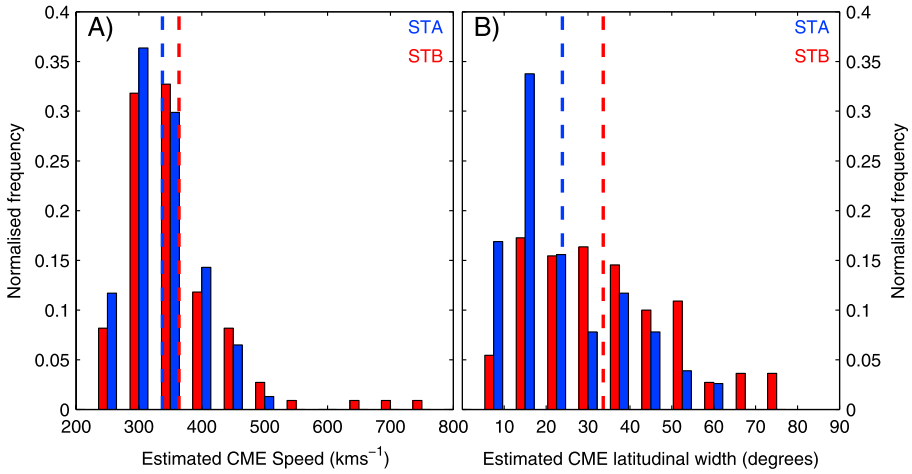

Figure 9. (a, b) Histograms of the distributions of the CME speeds and latitudinal widths, respectively, estimated from fitting the FPF fitting method to the t-ϵ profile along the central PA of each event. Events from STA are shown in red and events from STB are shown in blue. One point on the STA speed distribution lies outside the x axis limit at V_{cme} = 1330 km s^{-1}. The vertical dashed lines mark the means of each distribution. The speed and latitudinal width histograms use bin widths of 50 km s^{-1} and 7.5°, respectively.

one event from STB span the full range of available PAs, so we assume this is likely to have a small effect on the distributions presented in Figure 9b. The mean CME latitudinal width is 33.6±1.7° for STA and 23.9±1.6° for STB. There appears to be a systematic difference in the CME widths observed by STA, relative to STB, with STA typically reporting larger CME widths. Presently, the reason for this is not clear, but we suggest that this probably relates to the fact that the HI images from STB are frequently noisier than those from STA. This could make fainter transient profiles, as might be expected from the edge of a CME, harder to identify and track.

7. Summary

This study has detailed the construction of a catalogue of CMEs observed in HI images from STA and STB, made possible by the large database of CME t-ϵ profiles generated by the Solar Stormwatch citizen science project, in particular the Trace-it! activity. It is, as far as we know, the first CME catalogue generated as part of a citizen science project and one of the few catalogues that follows CMEs out to large elongations into the heliosphere. The duration of the catalogue extends from January 2007 to February 2010, over which time 110 CMEs were identified in the HI images from STA and 77 CMEs were identified in the HI images from STB.

The catalogue was produced by statistically reducing the raw set of t-ϵ profiles into clusters of profiles representing individual events, which were then averaged along each PA over which the event was observed, to provide a set of consensus profiles for each event. Approximately 30% of the Trace-it! data were used to create the catalogue, and so there is probably substantially more information on other solar wind features (including CMEs) within the Trace-it! data. Only 30% of the data were used due to the high threshold set to define the clusters of profiles associated with CMEs. This threshold was chosen to give confidence that CMEs have been robustly identified; clearly, this comes at the cost of not making fully efficient use of the data. Consequently, there is probably a bias in the Solar Stormwatch catalogue for including the biggest and brightest CMEs, with smaller, fainter, more marginal events not being identified.

So far, only approximately one half of the data from the HI instruments have been analyzed by Solar Stormwatch, which corresponds in particular to the period of minimum solar activity between solar cycles 23 and 24. The next stage of the project is to serve the Solar Stormwatch system the second half of the HI data, which spans the rise to maximum of solar cycle 24. In this period of increasing solar activity, we expect to find an increased occurrence frequency of CMEs, and it has been shown that this period contains a more numerous population of fast CMEs [*Möstl et al.*, 2014] than were found in this

©2014. The Authors.

study. As the CME speed is one of several important factors in determining the geo-effectiveness of Earth-impacting CMEs, inclusion of more energetic events will greatly enhance the usefulness of the Solar Stormwatch catalogue.

Additional future work will assess how the t-ϵ profiles manually identified by a single "expert-user" compare with the Solar Stormwatch consensus profiles and those of an automated algorithm of solar transient detection (L. Barnard et al., in preparation, 2014).

Citizen science projects allow us to achieve science goals which may otherwise seem unobtainable or impractical, while also allowing us to communicate the interesting science we research back to a wide and enthusiastic audience. The main aim of these projects is to facilitate academic research, and so we hope that the Solar Stormwatch CME catalogue finds use within the space weather community.

Appendix A: Additional Details of the Solar Stormwatch System and Activities

A1. The Zooniverse Platform

Behind the website, SSW is the third citizen science project created using the Zooniverse platform [*Fortson et al.*, 2011]. Built originally for Galaxy Zoo 2 [*Willett et al.*, 2013], the Zooniverse platform now supports more than 30 projects. It has undergone much development since SSW was built in 2009 but is designed primarily as a tool for serving up a large collection of "subjects" (for example, images or video) to an online interface and collecting back user-generated annotations of these assets in collections called "classifications." SSW is made up of a Ruby on Rails application backed by a MySQL database with user authentication handled by the Zooniverse login server. SSW volunteers can discuss the site and interesting images/video on the SSW Forum. As Zooniverse volunteers, they are part of a 1.2 million strong global community that can move between all Zooniverse projects, and many will receive regular email updates on project progress and other Zooniverse news. The Zooniverse has supported hundreds of millions of classifications since 2009 and branched beyond astronomy into multiple research domains.

A2. The Other Solar Stormwatch Activities

Below, we provide descriptions of the Solar Stormwatch activities that were not described in the main body of the text.

A2.1. Incoming!

The *Incoming!* activity mirrors the Spot! activity, but rather than serving a random movie file from archived science data, it displays movie pairs obtained from the near-real-time space weather beacon. In order to provide these data with as little latency as possible, differenced images are used to create the movies. The subsequent identification and tracking of CMEs with this SSW interface is more challenging owing to the reduced resolution and abstract nature of the differenced images and the frequent data gaps due to telemetry constraints. As more data become available, the real-time movies are updated to present the most recent data. Only after the data have been updated is each participant able to reclassify the CME.

A2.2. Incoming Trace-It!

Incoming trace-it! mirrors the Trace-it! activity in that a J-map, formed from the most recently available real-time beacon data, is served to the participant, who is required to identify the t-ϵ profile corresponding to the most recently occurring CME. Rather than being given a suggestion as to which profile to mark, the real-time data require the participant to instead scale the most recent feature that is observed. The J-maps are continually updated to present the most recently available data. Only after the J-map has been updated is each participant able to retrack the most recent CME. Recent work by K. Tucker-Hood et al. (submitted manuscript, 2014) uses the *Incoming trace-it!* data to investigate the efficacy of CME arrival predictions using the real-time HI observations.

A2.3. What's That?

In the *What's that?* activity, participants are asked to mark the frames of any movie in which they think there is something unusual. Four predetermined categories are presented: "comet," "dust impact," "optical effect," and "something else." A brief tutorial demonstrates the likely appearance of such features in the movies. Comets appear as features moving quickly relative to the star background, often with a prominent tail. Dust impacts on the spacecraft result in bright trails of secondary debris drifting through the HI field of view for no longer than a single frame. Optical effects, such as ghosting, result when a bright object is in or near the HI field of view, causing light to be scattered through the HI optics with sufficient intensity that it is comparable in brightness with the solar wind features HI was designed to track.

©2014. The Authors.

The fourth category is presented to identify frames in which there is something unexpected. Mostly these have resulted from glitches in the data processing or data blocks missing from the telemetry. Building a catalogue of such events is very useful in providing context to the quality of the CME observations and in achieving some secondary scientific goals such as tracking the distribution of dust around the STEREO spacecraft orbits.

A2.4. Track-It-Back

Track-it-back is the most complex of the activities within SSW. The participant is invited to take CMEs previously identified in the HI data and chart their initial progress from the solar disk and through the COR1 and COR2 instruments on either STA or STB. A movie is presented from each of these instruments. First, the participants are asked to estimate the speed of each CME by marking its first appearance and expansion across both of the COR1 and COR2 movies. In addition, the participant is also asked to mark the angular extent of the CME in order to determine its PA direction and to estimate any possible relationship between the speed and angular extent of CMEs in the STEREO data set. The participant is then shown images of the solar disk made using the four extreme ultraviolet wavelengths from the EUVI instrument and asked to choose the wavelength that reveals the erupting material most clearly. This gives an indication of the temperature of the erupting material. Once this has been selected, the participant is then asked to mark the location of the source of the eruption on the solar disk.

Appendix B: Profile Clustering Using a Fixed Window Around Spot! CME Times

Initially, it seemed sensible to group the t-ϵ profiles into individual CMEs by comparison with the CME times identified in Spot!. Over the period analyzed, the use of Spot! resulted in the identification of 145 CMEs in STA and 113 in STB, 50 of which were seen by both spacecraft. To associate the t-ϵ profiles with the Spot! CME times, we searched for every t-ϵ profile that began within a fixed window of the Spot! CME start time. It was found that using a window size of $\pm$18 h (comparable with the transit time of a slow CME across the HI1 FOV) is reasonably effective at associating t-ϵ profiles with the Spot! CMEs. However, there were a set of instances where the Spot! CMEs had very few t-ϵ profiles associated with them, and even none in some instances. We decided to investigate the reason for this by considering all of the Spot! CMEs for which less than 5 t-ϵ profiles could be found within $\pm$18 h of the CME time of appearance in the HI1 FOV. There were 16 events that matched these criteria, 11 from STA and 5 from STB. For each of these cases, we inspected the relevant HI1 differenced image movie and J-map. In four cases we could not visibly identify a CME trace in the corresponding HI1 movie, but there were optical artifacts in the HI1 images that could have been mistakenly identified as a CME. In these instances, there was no CME profile to identify in the corresponding J-maps, and it is therefore both appropriate and reassuring that no sensible profiles were generated by the participants. In one instance, a data gap approximately 1 day after the event onset meant that the t-ϵ profile was very poorly defined in the J-map. In 8 of these 16 events, we were able to identify narrow outflows of plasma that might not be considered as CMEs. Inspecting the J-maps revealed that these events were not visible over more than 20° in PA, and six of them faded below the brightness noise floor within the HI1 FOV. These were therefore insubstantial events, which left traces that were difficult to identify in the corresponding J-maps. The remaining two events were due to a single well-defined CME, seen in both STA and STB, which occurred approximately 1 day before the end of the very final Trace-it! J-maps; consequently, the t-ϵ profile was not clear and could have only been followed over a short elongation range. Therefore, the poor associations were due primarily to CMEs leaving poor traces in the J-maps and also the incorrect identification of optical artifacts in the images.

As part of this investigation, we also inspected subsets of the t-ϵ profiles overlaid on the corresponding J-maps, along with the Spot! CME times. This revealed that there were also well-defined events that had been identified in the Trace-it! activity, which had not been identified by Spot!. This can happen if a short data gap in the HI1 movies obscures the launch of a CME. While this makes the CME impossible to identify in the HI1 movies, it can still be identified in the HI J-maps. Therefore, to maximize our use of the Trace-it! data, we decided that in preference to searching for profiles within a fixed window of the Spot! CMEs, we would instead search for clusters in the t-ϵ profiles to yield a catalogue of CMEs that were visible in the HI J-maps. We do, however, note that this does not make the results here independent of the Spot! results, as the J-maps analyzed in Trace-it! are driven by the Spot! results.

©2014. The Authors.

Acknowledgments
The Solar Stormwatch CME catalogue and the raw Solar Stormwatch data used in its construction are available to download at www.met.reading.ac.uk/~spate/solarstormwatch. The STEREO/HI data are available for download from http://www.ukssdc.rl.ac.uk/solar/stereo/data.html. We thank the UK's Natural Environmental Research Council (NERC) for support under grant NE/J024678/1. Solar Stormwatch is a joint project between the Royal Observatory Greenwich, the Rutherford Appleton Laboratory Space division, the Zooniverse team, and the University of Reading. STEREO is the third mission in NASA's Solar Terrestrial Probes program. STEREO/HI was developed by a consortium comprising RAL, the University of Birmingham (UK), CSL (Belgium), and NRL (USA). This publication has been made possible by the participation of more than 16,000 volunteers in the Solar Stormwatch project (http://www.solarstormwatch.com/authors). This publication has been made possible by the participation of more than 16,000 volunteers in the Solar Stormwatch project (http://www.solarstormwatch.com/authors).

References

Barnard, L., and M. Lockwood (2011), A survey of gradual solar energetic particle events, *J. Geophys. Res.*, *116*, A05103, doi:10.1029/2010JA016133.

Borovsky, J. E., and M. H. Denton (2006), Differences between CME-driven storms and CIR-driven storms, *J. Geophys. Res.*, *111*, A07S08, doi:10.1029/2005JA011447.

Brueckner, G. E., et al. (1995), The Large Angle Spectroscopic Coronagraph (LASCO), *Sol. Phys.*, *162*(1–2), 357–402, doi:10.1007/BF00733434.

Byrne, J. P., H. Morgan, S. R. Habbal, and P. T. Gallagher (2012), Automatic detection and tracking of coronal mass ejections. II. Multiscale filtering of coronagraph images, *Astrophys. J.*, *752*(2), 145, doi:10.1088/0004-637X/752/2/145.

Cannon, P., et al. (2013), Extreme space weather: Impacts on engineered systems and infrastructure, *Tech. Rep.*, Royal Academy of Engineering, London.

Davies, J. A., R. A. Harrison, A. P. Rouillard, N. R. Sheeley, C. H. Perry, D. Bewsher, C. J. Davis, C. J. Eyles, S. R. Crothers, and D. S. Brown (2009), A synoptic view of solar transient evolution in the inner heliosphere using the Heliospheric Imagers on STEREO, *Geophys. Res. Lett.*, *36*, L02102, doi:10.1029/2008GL036182.

Davies, J. A., et al. (2012), A self-similar expansion model for use in solar wind transient propagation studies, *Astrophys. J.*, *750*(1), 23, doi:10.1088/0004-637X/750/1/23.

Davis, C. J., J. A. Davies, M. Lockwood, A. P. Rouillard, C. J. Eyles, and R. A. Harrison (2009), Stereoscopic imaging of an Earth-impacting solar coronal mass ejection: A major milestone for the STEREO mission, *Geophys. Res. Lett.*, *36*, L08102, doi:10.1029/2009GL038021.

Davis, C. J., J. Kennedy, and J. A. Davies (2010), Assessing the accuracy of CME speed and trajectory estimates from STEREO observations through a comparison of independent methods, *Sol. Phys.*, *263*(1–2), 209–222, doi:10.1007/s11207-010-9535-2.

Davis, C. J., et al. (2011), A comparison of space weather analysis techniques used to predict the arrival of the Earth-directed CME and its shockwave launched on 8 April 2010, *Space Weather*, *9*, S01005, doi:10.1029/2010SW000620.

Davis, C. J., et al. (2012a), The distribution of interplanetary dust between 0.96 and 1.04 au as inferred from impacts on the STEREO spacecraft observed by the heliospheric imagers, *Mon. Not. R. Astron. Soc.*, *420*(2), 1355–1366, doi:10.1111/j.1365-2966.2011.20125.x.

Davis, C. J., J. A. Davies, M. J. Owens, and M. Lockwood (2012b), Predicting the arrival of high-speed solar wind streams at Earth using the STEREO Heliospheric Imagers, *Space Weather*, *10*, S02003, doi:10.1029/2011SW000737.

Eyles, C. J., et al. (2008), The heliospheric imagers onboard the STEREO mission, *Sol. Phys.*, *254*(2), 387–445, doi:10.1007/s11207-008-9299-0.

Floyd, O., P. Lamy, Y. Boursier, and A. Llebaria (2013), ARTEMIS. II: A second-generation catalog of LASCO coronal mass ejections including mass and kinetic energy, *Sol. Phys.*, *288*(1), 269–289, doi:10.1007/s11207-013-0281-0.

Fortson, L., K. Masters, R. Nichol, K. D. Borne, E. M. Edmondson, C. Lintott, J. Raddick, K. Schawinski, and J. Wallin (2011), *Galaxy Zoo: Morphological Classification and Citizen Science, Advances in Machine Learning and Data Mining for Astronomy*, CRC Press, Taylor and Frances Group, Boca Raton, Fla.

Gonzalez, W. D., A. L. Clúa de Gonzalez, J. H. A. Sobral, A. Dal Lago, and L. E. A. Vieira (2001), Solar and interplanetary causes of very intense geomagnetic storms, *J. Atmos. Sol. Terr. Phys.*, *63*(5), 403–412, doi:10.1016/S1364-6826(00)00168-1.

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

Hapgood, M. A. (2011), Towards a scientific understanding of the risk from extreme space weather, *Adv. Space Res.*, *47*(12), 2059–2072, doi:10.1016/j.asr.2010.02.007.

Holley, R. (2009), *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*, National Library of Australia, Canberra.

Howard, R. A., et al. (2008), Sun Earth connection coronal and heliospheric investigation (SECCHI), *Space Sci. Rev.*, *136*(1–4), 67–115, doi:10.1007/s11214-008-9341-4.

Lugaz, N. (2010), Accuracy and limitations of fitting and stereoscopic methods to determine the direction of coronal mass ejections from heliospheric imagers observations, *Sol. Phys.*, *267*(2), 411–429, doi:10.1007/s11207-010-9654-9.

Morgan, H., J. P. Byrne, and S. R. Habbal (2012), Automatically detecting and tracking coronal mass ejections. I. Separation of dynamic and quiescent components in coronagraph images, *Astrophys. J.*, *752*(2), 144, doi:10.1088/0004-637X/752/2/144.

Möstl, C., and J. A. Davies (2012), Speeds and arrival times of solar transients approximated by self-similar expanding circular fronts, *Sol. Phys.*, *285*(1–2), 411–423, doi:10.1007/s11207-012-9978-8.

Möstl, C., et al. (2014), Connecting speeds, directions and arrival times of 22 coronal mass ejections from the Sun to 1AU, *Astrophys. J.*, *787*(2), 119, doi:10.1088/0004-637X/787/2/119.

National Research Council–Space Science Board (2008), *Severe Space Weather Events–Understanding Societal and Economic Impacts: A Workshop Report*, The National Academies Press, Washington, D. C.

Olmedo, O., J. Zhang, H. Wechsler, A. Poland, and K. Borne (2008), Automatic detection and tracking of coronal mass ejections in coronagraph time series, *Sol. Phys.*, *248*(2), 485–499, doi:10.1007/s11207-007-9104-5.

Raddick, M. J., G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenberg (2010), Galaxy zoo: Exploring the motivations of citizen science volunteers, *Astron. Educ. Rev.*, *9*(1), doi:10.3847/AER2009036.

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

Robbrecht, E., D. Berghmans, and R. A. M. Van der Linden (2009), Automated LASCO CME catalog for solar cycle 23: Are CME's scale invariant?, *Astrophys. J.*, *691*(2), 1222–1234, doi:10.1088/0004-637X/691/2/1222.

Rouillard, A. P., et al. (2008), First imaging of corotating interaction regions using the STEREO spacecraft, *Geophys. Res. Lett.*, *35*, L10110, doi:10.1029/2008GL033767.

Rouillard, A. P., et al. (2009), A solar storm observed from the Sun to Venus using the STEREO, Venus Express, and MESSENGER spacecraft, *J. Geophys. Res.*, *114*, A07106, doi:10.1029/2008JA014034.

Rouillard, A. P., et al. (2010), Intermittent release of transients in the slow solar wind: 1. Remote sensing observations, *J. Geophys. Res.*, *115*, A04103, doi:10.1029/2009JA014471.

Savani, N. P., et al. (2012), Observational tracking of the 2D structure of coronal mass ejections between the sun and 1 AU, *Sol. Phys.*, *279*(2), 517–535, doi:10.1007/s11207-012-0041-6.

Sheeley, N. R., J. H. Walters, Y.-M. Wang, and R. A. Howard (1999), Continuous tracking of coronal outflows: Two kinds of coronal mass ejections, *J. Geophys. Res.*, *104*(A11), 24,739–24,767, doi:10.1029/1999JA900308.

Sheeley, N. R., Jr., et al. (2008), Heliospheric images of the solar wind at Earth, *Astrophys. J.*, *675*(1), 853–862, doi:10.1086/526422.

Tappin, S. J., T. A. Howard, M. M. Hampson, R. N. Thompson, and C. E. Burns (2012), On the autonomous detection of coronal mass ejections in heliospheric imager data, *J. Geophys. Res.*, *117*, A05103, doi:10.1029/2011JA017439.

©2014. The Authors.

UK Cabinet Office (2013), National Risk Register of Civil Emergencies 2013, *Tech. Rep.*, United Kingdom's Cabinet Office, London.
Webb, D. F., and T. A. Howard (2012), Coronal mass ejections: observations, *Living Rev. Sol. Phys.*, *9*, 3.
Webb, D. F., et al. (2006), Solar Mass Ejection Imager (SMEI) observations of coronal mass ejections (CMEs) in the heliosphere, *J. Geophys. Res.*, *111*, A12101, doi:10.1029/2006JA011655.
Willett, K. W., et al. (2013), Galaxy zoo 2: Detailed morphological classifications for 304122 galaxies from the Sloan Digital Sky Survey, *Mon. Not. R. Astron. Soc.*, *435*, 2835–2860, doi:10.1093/mnras/stt1458.
Williams, A. O., J. A. Davies, S. E. Milan, A. P. Rouillard, C. J. Davis, C. H. Perry, and R. A. Harrison (2009), Deriving solar transient characteristics from single spacecraft STEREO/HI elongation variations: A theoretical assessment of the technique, *Ann. Geophys.*, *27*(12), 4359–4368, doi:10.5194/angeo-27-4359-2009.
Yashiro, S., G. Michalek, and N. Gopalswamy (2008), A comparison of coronal mass ejections identified by manual and automatic methods, *Ann. Geophys.*, *26*(10), 3103–3112, doi:10.5194/angeo-26-3103-2008.

©2014. The Authors.