submitted to **Journal of Space Weather and Space Climate**

Investigating the Kinematics of Coronal Mass Ejections with the Automated CORIMP Catalog

Jason P. Byrne

Rutherford Appleton Laboratory, STFC: RAL Space, Harwell Oxford, OX11 0QX, UK.
e-mail: jason.byrne@stfc.ac.uk

ABSTRACT

Studying coronal mass ejections (CMEs) in coronagraph data can be challenging due to their diffuse structure and transient nature, compounded by the variations in their dynamics, morphology, and frequency of occurrence. The large amounts of data available from missions like SOHO makes manual cataloging of CMEs tedious and prone to human error, and so a robust method of detection and analysis is required and often preferred. This has led to the development of automated CME detection and tracking algorithms such as those implemented in the CACTus and SEEDS catalogs. However, drawbacks do exist: for example, the CACTus method of detection fails to resolve CME acceleration profiles; the SEEDS method of detection is limited to the LASCO/C2 field-of-view; and both CACTus and SEEDS running-difference images suffer from spatiotemporal crosstalk. The new CORIMP catalog was developed in an effort to overcome these drawbacks, through the implementation of a dynamic background separation technique and multiscale edge detection, that together isolate and characterise CME structure throughout the LASCO field-of-view. CORIMP also applies a Savitzky-Golay filter, along with quadratic and linear fits, to the height-time measurements for better revealing the true CME velocity and acceleration profiles across the plane-of-sky. Here we present a sample of new results from the CORIMP CME catalog, and directly compare them with the other automated CACTus and SEEDS catalogs, as well as the manual CDAW catalog and a previous study of the sample events. We further investigate a form of unsupervised machine learning by means of a *k*-means clustering algorithm to distinguish detections of multiple CMEs that occur close together in space and time. While challenges still exist, this investigation and comparison of results demonstrates the reliability and robustness of the CORIMP catalog, proving its efficacy in detecting and tracking CMEs throughout the LASCO dataset.

Key words. Sun – Coronal Mass Ejection (CME) – Space weather – Solar image processing – Machine learning

1. Introduction

Coronal mass ejections (CMEs) represent the largest, most dynamic phenomena that originate from the Sun (Chen, 2011; Webb and Howard, 2012). Propagating at speeds of up to thousands of kilometres per second, with energies on the order of 10^{32} ergs, they can drive adverse space weather throughout the solar system (Howard and Tappin, 2005; Pulkkinen, 2007). Given their potentially hazardous impact on Earth's geomagnetic environment, the physics governing their eruption and propagation needs to be understood so that their effects may be predicted in the guise of space-weather forecasting. To this end, observations of CMEs must be rigorously inspected in order to determine their dynamics, and this is most generally undertaken with the use of coronagraph instruments (e.g., Koomen et al., 1975; Sheeley et al., 1980; MacQueen et al., 1980; Illing and Hundhausen, 1985; Hundhausen, 1993; Brueckner et al., 1995; Howard et al., 2008).

CMEs tend to be faint, transient phenomena, observed in white-light images that are prone to noise and user-dependent bias in their interpretation. During solar minimum they can occur every few days, but at solar maximum there can be several per day (St. Cyr et al., 2000; Yashiro et al., 2004). They exhibit a wide variety of morphologies, moving in unpredictable directions and speeds in the solar wind, with differing levels of possible geo-effectiveness (Plunkett et al., 2001; Schwenn et al., 2005; Davis et al., 2009; Lugaz and Kintner, 2012). A wealth of image processing techniques have been explored to study CMEs in remote-sensing image data provided by such instruments as the Large Angle Spectroscopic Coronagraph (LASCO; Brueckner et al., 1995) onboard the Solar and Heliospheric Observatory (SOHO; Domingo et al., 1995). These techniques generally rely on some form of image differencing to highlight moving features in the observed intensities, but this introduces spatiotemporal crosstalk and scaling issues that affect the accuracy of CME characterisations. More advanced image processing methods have thus been explored, such as optical flow techniques (Colaninno and Vourlidas, 2006), supervised segmentation techniques (Goussies et al., 2010), wavelets (Stenborg and Cobelli, 2003) and curvelets (Gallagher et al., 2011). The large volume of data available has made it necessary to automate such techniques for detecting and tracking CMEs across images, with a view to cataloging their kinematics and morphologies. This allows for more robust CME detections by avoiding the troublesome effects of standard image differencing techniques. It is therefore possible to maintain a non-biased characterisation of the CME structure in every event, since automated techniques are self-consistent.

To date, a point-and-click catalog of CMEs in LASCO data has been undertaken at the Coordinated Data Analysis Workshop (CDAW) Data Center (Gopalswamy et al., 2009), which operates by tracking CMEs in running difference images to produce information on the dynamics of each event. It is a wholly manual procedure and is therefore subject to user bias in interpreting the data. Automated catalogs have since been developed to overcome this bias and tedium. The Computer Aided CME Tracking routine (CACTus; Robbrecht and Berghmans, 2004) is the first such automated catalog, that works by using a Hough transform to detect intensity ridges corresponding to CME tracks in time-height stacks (J-maps) of polar-unwrapped running-difference LASCO images. The Solar Eruptive Events Detection System (SEEDS; Olmedo et al., 2008) is another automated catalog that similarly uses polar-unwrapped running-difference LASCO images but with a form of threshold segmentation to approximate the shape of the CME leading edge in every image. A new automated catalog has recently been developed from a unique set of coronal image processing techniques, called CORIMP, that overcomes many of the limitations of current catalogs in operation (Morgan et al., 2012; Byrne et al., 2012). An online database has been produced for the SOHO/LASCO data and event detections therein; providing information on CME onset time, position angle, angular width, speed, acceleration, and mass, along with kinematic plots and observation movies. By investigating the catalog output it is intended that this work will lead to an improved understanding of the dynamics of CMEs. Furthermore, a realtime version of the algorithm has been implemented to provide CME detection alerts to the interested space weather community.

In Sect. 2 the CME catalogs are discussed in greater detail to highlight their methodologies and drawbacks. A sample of CMEs is then investigated in Sect. 3 in order to compare the outputs of each catalog, paying particular interest to the robustness and reliability of the new CORIMP catalog. In Sect. 4 a first effort is made to use a form of unsupervised machine learning to isolate spatially and temporally overlapping CME detections. The conclusions of this investigation are presented in Sect. 5.

2. Cataloging CMEs

In coronagraph images CMEs are observed as outwardly moving regions of stronger brightness intensities than the background corona (see the example in Fig. 1). Different methods for thresholding the CME intensity in such images have been employed by different catalogs. This is in order to detect their appearance and track their motion through the field-of-view, leading to a determination of the CME kinematics and morphology. However, each method suffers from drawbacks and, as such, the resulting CME catalogs can vary significantly in their characterisations and measurements of each event.

2.1. CORIMP Automated Catalog

The CORIMP[1] catalog was developed with a method of dynamic signal separation and multiscale edge detection to overcome certain drawbacks of previous detection and tracking methods that rely on running-difference images. The CME signal is separated from the more quiescent streamers and coronal structures in LASCO/C2 and C3 images, such that a multiscale filtering technique may be used to suppress noise in order to characterise the CME structure and track its motion. A spread of heights is then measured across the angular span of the CME. A type of cleaning algorithm is applied to the height-time measurements before the kinematics are determined. This is required to overcome cases where pixels along the CME front edges at a specific position angle are not correctly identified, but rather a pixel corresponding to core material behind the CME front is measured, which causes unnecessary scatter in the height-time datapoints. The cleaning algorithm works by stepping along the height-time measurements at each position angle and requiring that only sequentially increasing heights are plotted. This helps remove the detections of trailing CME material from the height-time plots such that only the kinematics of the main CME front are determined. The kinematics are then derived in three ways: using a Savitzky-Golar filter (similar to a spline), a quadratic fit (second-order), and a linear fit (first-order/straight-line). The importance of a robust method for determining the kinematics of a transient event is discussed in Byrne et al. (2013), wherein the often-used method of 3-point Lagrangian interpolation and associated error-propagation were shown to behave counter-intuitively and provide misleading kinematic results. Given the large variety of CMEs that can occur with differing kinematic profiles, the inspection of these three automatic fitting techniques in CORIMP can better reveal the true trends of CME motion.

[1] http://alshamess.ifa.hawaii.edu/CORIMP/

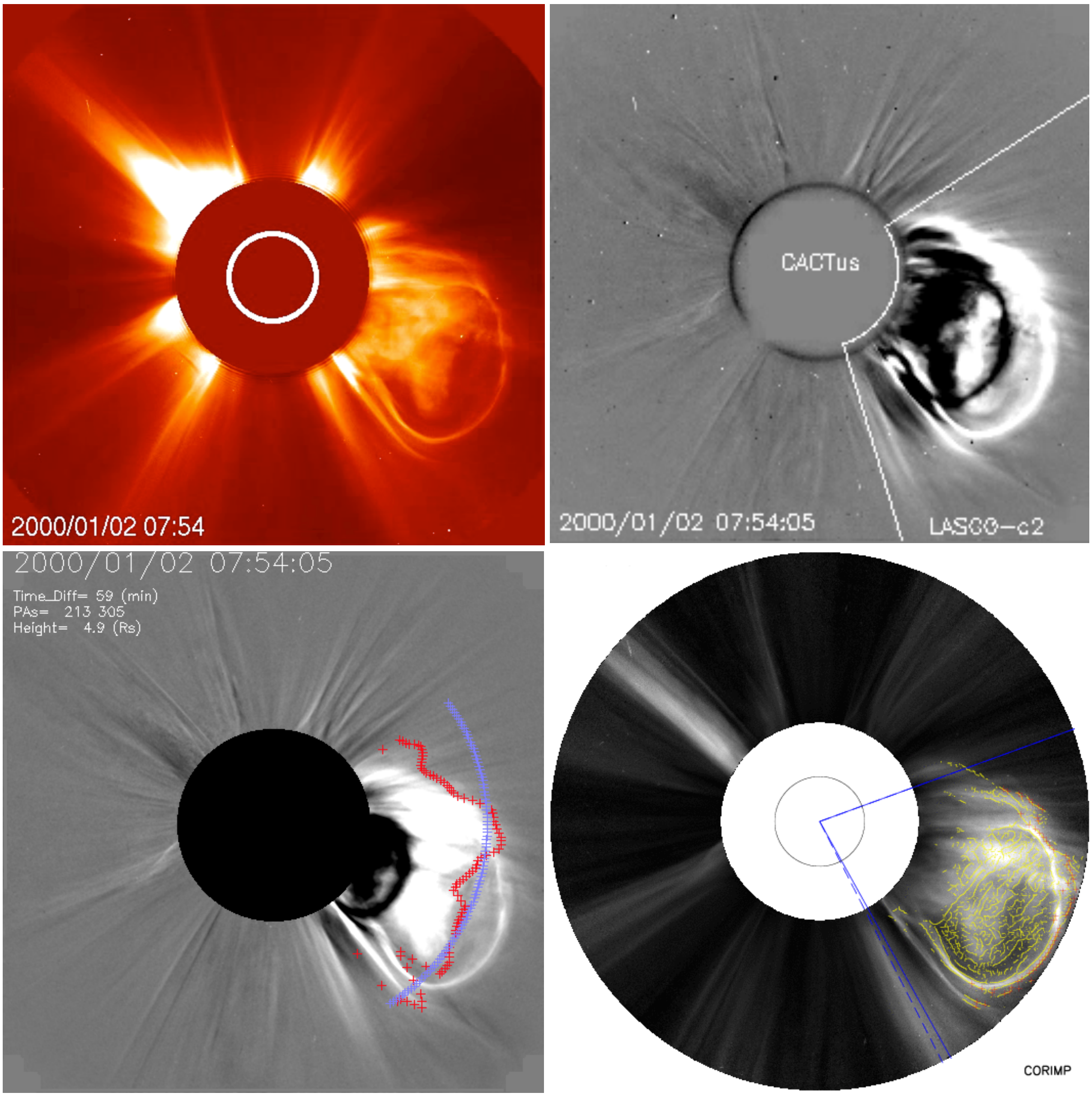

Fig. 1. LASCO/C2 observations of a CME on 2000 Jan. 02 at 07:54 UT. *Top left:* Level 2 processed image reproduced from the CDAW catalog. *Top right:* Running difference image reproduced from the CACTus catalog with the angular span of the CME detection indicated in white. *Bottom left:* Running difference image reproduced from the SEEDS catalog with the CME front detection highlighted in red (and the extended 'half-max lead' in purple). *Bottom right:* Normalised radial-gradient filtered (NRGF) image taken from the CORIMP catalog with the angular span of the CME detection indicated in blue, the pixel-chained CME structure in yellow, and the CME front in red.

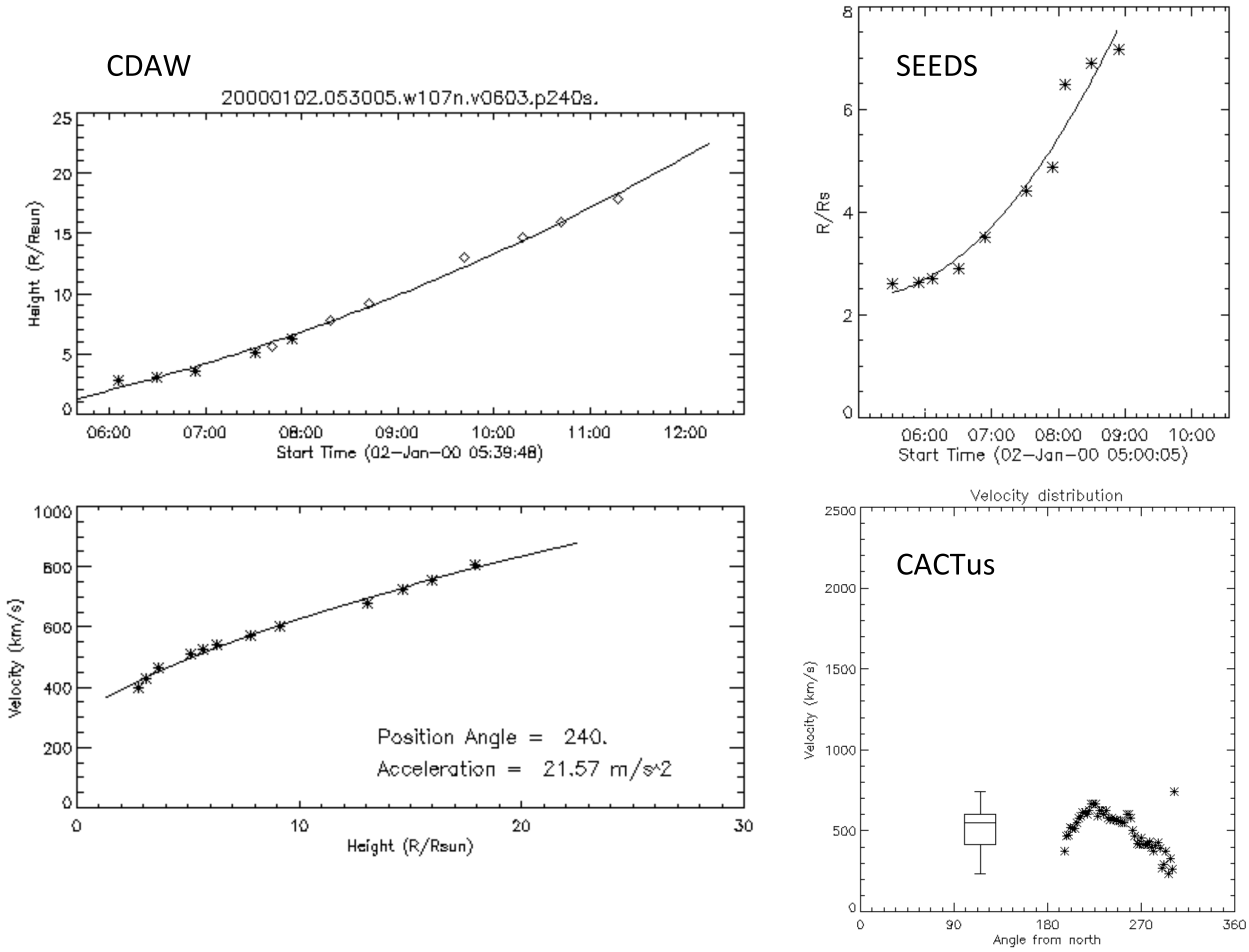

Fig. 2. The kinematic outputs for the 2000 Jan. 02 CME reproduced from the CDAW, SEEDS and CACTus catalogs. *Top left:* The CDAW catalog height-time measurements of the CME chosen manually along the running-difference bright front (at position angle 240) with a second-order fit. *Bottom left:* The corresponding CDAW velocity profile plotted against height, showing an acceleration of 21.57 $m\,s^{-2}$. *Top right:* The automated SEEDS height-time measurements and second-order fit resulting in an acceleration of 18.6 $m\,s^{-2}$, in the LASCO/C2 field-of-view. *Bottom right:* The automated CACTus velocities determined along the angular span of the CME, with a corresponding box-and-whisker plot to highlight the median (548 $km\,s^{-1}$) and interquartile range.

2.2. CACTus Automated Catalog

The CACTus[2] catalog was the first automated CME detection algorithm, in operation since 2004. It is based upon the detection of CMEs as bright ridges in time-height slices at each angle around a coronagraph image. A running-difference technique is applied and each image is transformed into Sun-centred polar coordinates, re-binned, and the C2 and C3 fields-of-view are combined. These are then stacked in time, and for each angle the corresponding time-height slice undergoes a modified Hough transform for detecting intensity ridges across it. Thresholding the most significant ridges

[2] http://sidc.oma.be/cactus/

filters out the progression of CMEs, with the variables for each ridge characterised by onset time, velocity, and position angle. A median velocity across the angular span of each event is quoted as the CME speed.

The running-difference cadence, the ridge intensity threshold, and the imposed limit on how many frames a CME may exist all affect how successful the automated detection can be. However, Robbrecht and Berghmans (2004) show the algorithm to be robust in reproducing the detections of a human user by directly comparing with the CDAW catalog. The main drawback of the CACTus catalog for studying CMEs is the imposed zero acceleration of the detection algorithm, since the Hough transform thresholds the ridges as straight lines whose slopes provide a constant velocity. The velocity itself may also be an underestimate since it is a median across the span of the CME (see the example velocity-plot at the bottom right of Fig. 2). The angular spans are possibly over-estimated since side outflows in the images are enhanced by the running-difference and may include streamer deflections. It is also difficult to distinguish when one CME has fully progressed from the field-of-view and another CME has entered it, so in some cases trailing portions of a CME are detected as separate events.

2.3. SEEDS Automated Catalog

The SEEDS[3] catalog employs an automated CME detection algorithm for tracking an intensity thresholded CME front in running-difference images from LASCO/C2. The images are unwrapped into Sun-centred polar coordinates, and a normalised running-difference technique is applied (such that the mean intensity of the new image is effectively zero). The pixel intensities (positive values only) are then summed along angles and thresholded at a certain number of standard deviations above the mean intensity. This determines the "core angles" of the CME, and a region growing technique based on a secondary threshold of intensities in the rest of the image is applied to open the angular span to include the full CME. An issue arises when streamer deflections occur that offset the region growing technique and overestimate the CME angular width. An intensity average across the angles within the span of the CME is then determined, and where the forward portion of this intensity profile equals half its maximum value is taken as the CME height. The velocity and acceleration are determined from the heights through consecutive images and these results are output with the CME position angle and angular width in the SEEDS catalog (see the example height-time plot at the top right of Fig. 2).

Along with the issues of streamer deflections and the tracking being limited to only the C2 field-of-view, the choice of the "Half-Max-Lead" as the CME height is dependant on the overall CME brightness, and thus any brightness change during its propagation will affect this measurement. This adds to the error on the height-time profile, which affects the accuracy of the derived velocity and acceleration.

2.4. CDAW Manual Catalog

The CME catalog hosted at the CDAW Data Center[4] grew out of a necessity to record a simple but effective description and analysis of each event observed with LASCO (Gopalswamy et al.,

[3] http://spaceweather.gmu.edu/seeds/
[4] http://cdaw.gsfc.nasa.gov/CME_list

2009). The catalog is wholly manual in its operation, with a user tracking the CME through C2 and C3 running-difference images and producing a "point-&-click" height-time plot of each event. A linear fit to the height-time profiles provides a 1st-order estimate for the plane-of-sky velocity, and a quadratic fit provides a 2nd-order velocity fit and an acceleration for the event. The central position angle and angular width of the CME are also deduced from the images, and the event is flagged as a halo if it spans 360°, partial halo if it spans $\geq 120°$, and wide if it spans $\geq 60°$. The catalog itself lists each CME's first appearance in C2, central position angle, angular width, linear speed, 2nd-order speed at final height, 2nd-order speed at 20 $R_\odot$, acceleration, mass, kinetic energy, and measurement position angle (the angle along which the heights of the CME are determined; see the example height-time and velocity-height plots in the left of Fig. 2). While the human eye is supremely effective at distinguishing CMEs in coronagraph images, errors may be introduced to the manual cataloging procedure through the biases of different operators; for example, in deciding how the images are scaled, where along the CME the heights are measured, or whether a CME is worth including in, or discarding from, the catalog.

3. CME Event Sample and Catalog Results

A selection of CMEs from the SOHO/LASCO data was chosen in the analysis of Byrne et al. (2009), wherein multiscale methods of edge detection and a resulting ellipse characterisation of the CME front were used to track its apex. These events were chosen based on their varying styles of eruption and appearance, in order to compare with the measurements of the manual CDAW catalog (see images of each CME in Fig. 5 of Byrne et al. 2009). The images were not differenced to avoid spatiotemporal cross-talk and associated scaling issues. The uncertainties on the height measurements were quantified by the multiscale filter size and subsequent ellipse-fitting, and propagated into the kinematics via numerical differentiation using 3-point Lagrangian interpolation. However, it has since been demonstrated by Byrne et al. (2013) that this method for deriving kinematics is not wholly reliable, and other approaches must be considered as discussed in Sect. 2.1 above. Following the development of the automated CORIMP algorithms, these events are now revisited in this new catalog and directly compared with the other automated CACTus and SEEDS catalogs, the manual CDAW catalog, and the results of Byrne et al. (2009). In each case below, the tabled information and kinematic plots are reproduced directly from the online catalogs, and not rescaled or otherwise manipulated, for a fair comparison. Note, in the tables some values have a corresponding maximum and/or minimum (x_{min}^{max}) as specified in the respective catalogs.

3.1. Arcade eruption: 2000 January 2

The CME that erupted off the southeast limb of the Sun on 2000 Jan. 02 from ~06:06 UT in LASCO exhibited an arcade-type structure consisting of multiple bright loops. CORIMP identified the bulk of the CME through the LASCO field-of-view to ~24 $R_\odot$. However, this CME may be deemed the third in a series of four CMEs that occurred in succession off the southeast limb, that CORIMP failed to separate due to their spatial and temporal overlap (essentially a smaller CME in between two large ones connects their detections along with a fourth smaller one afterwards, as seen in the CORIMP CME detection stacks reproduced in Fig. 3 - see Byrne et al. 2012 for details). This therefore serves as an example of the need to inspect the catalog output before trusting the quoted values

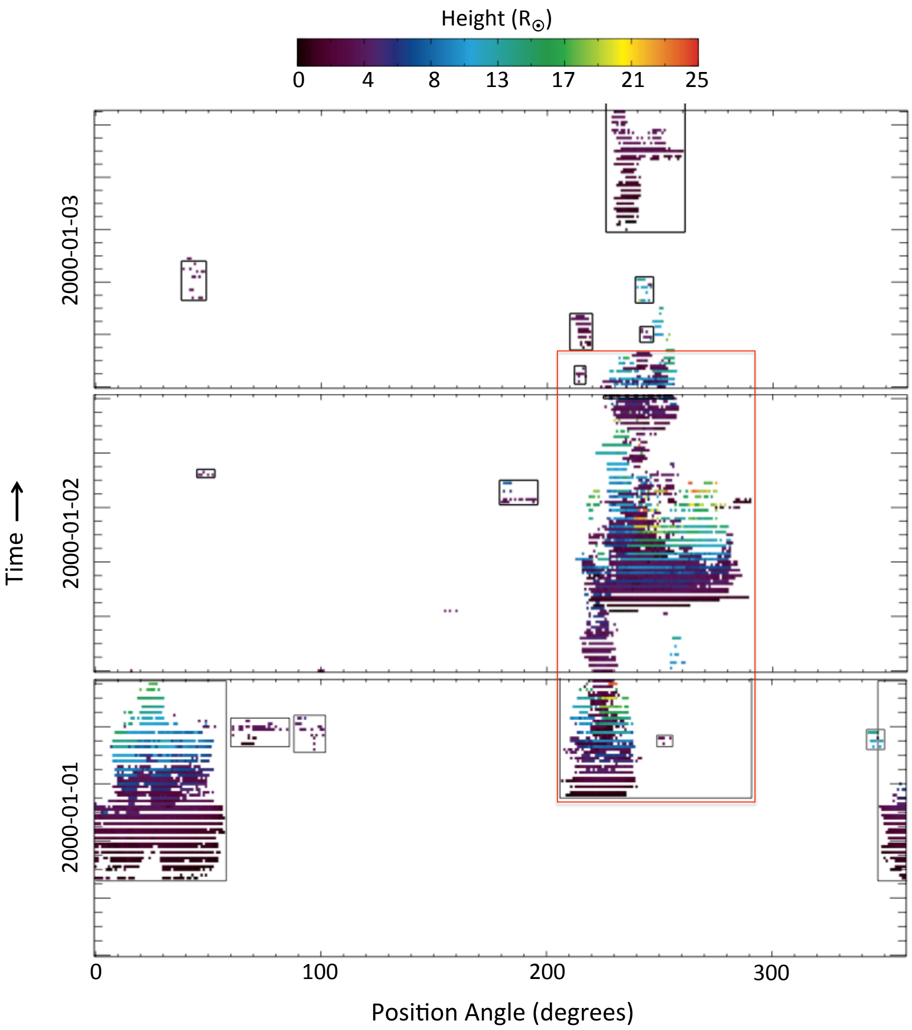

Fig. 3. Daily detection stacks reproduced from the CORIMP CME catalog for the SOHO/LASCO observations in the date range 2000 Jan. 01 – 03. These stacks are generated from the automatically measured CME front heights at every position angle in an image, stacked in time. CMEs appear as groupings of colour-graded pixels, as indicated by the boxed regions. The overlapping CMEs in this time interval are boxed in red, corresponding to the height-time profiles in Fig. 4 (that have been put through the "cleaning algorithm").

listed in Table 1. The CORIMP height-time measurements (in the time range ~06:00 – 16:00 UT in Fig. 4) reveal a non-linear trend indicative of an early acceleration that the Savitzky-Golay filter determines decreases from a maximum of ~$50\,m\,s^{-2}$ to $0\,m\,s^{-2}$, as the maximum velocity levels off in the range of ~$500 – 600\,km\,s^{-1}$. This is consistent with the measurements of Byrne et al. (2009) shown in their Fig. 6. The quadratic (and linear) fits in CORIMP agree with a maximum velocity in this range of ~$500 – 600\,km\,s^{-1}$ and an acceleration in the range of approximately $\pm 20\,m\,s^{-2}$. CACTus determined a linear velocity of $548\,km\,s^{-1}$ (in the range $231 – 744\,km\,s^{-1}$). SEEDS determined a linear velocity of $292\,km\,s^{-1}$ and an acceleration of $18.6\,m\,s^{-2}$ (in the C2 field-of-view). And CDAW determined a linear velocity of $603\,km\,s^{-1}$ and an overall acceleration of $21.6\,m\,s^{-2}$.

Catalog	CPA [*deg.*]	AW [*deg.*]	Lin. Speed [$km\,s^{-1}$]	Accel. [$m\,s^{-2}$]
CORIMP	250	81^{83}	454^{743}	1^{14}_{-17}
CACTus	250	106	548^{744}_{231}	
SEEDS	257	96	292	18.6
CDAW	253	107	603	21.6

Table 1. Catalog measurements of a CME observed by SOHO/LASCO on 2000 Jan. 02 from ~06:06 UT. Note, some values have a corresponding maximum and/or minimum (x^{max}_{min}) as specified in the respective catalogs.

These catalog measurements are listed in Table 1. (Note that the slightly lower angular width in CORIMP is due to the exclusion of part of the questionable streamer deflection/interaction along the southern flank of the CME.) Therefore, by inspection, the results of the CORIMP CME catalog are in relative agreement with the other catalogs and manual analysis, and CORIMP is deemed robust albeit unreliable at separating overlapping events.

3.2. Gradual/expanding CME: 2000 April 18

Catalog	CPA [*deg.*]	AW [*deg.*]	Lin. Speed [$km\,s^{-1}$]	Accel. [$m\,s^{-2}$]
CORIMP	210	98	431^{537}	4^{15}_{-11}
CACTus	198	102	463^{744}_{227}	
SEEDS	195	108	338	17.7
CDAW	195	105	668	23.1

Table 2. Catalog measurements of a CME observed by SOHO/LASCO on 2000 Apr. 18 from ~14:54 UT.

The CME that erupted off the south limb of the Sun on 2000 Apr. 18 from ~14:54 UT in LASCO exhibited a typical 3-part structure of leading CME front, cavity and bright core. CORIMP identified the bulk of the CME through the LASCO field-of-view to ~25 $R_\odot$, though it did not detect a southern portion of the faint CME front in the latter C3 observations. A western portion of material also erupted as a delayed part of the northern flank of the CME, that appears as a somewhat secondary height-time profile in the CORIMP kinematic plots in Fig. 5 (at position angles ~250° in the redder end of the colorbar). The CORIMP height-time measurements reveal a non-linear trend indicative of an early acceleration that the Savitzky-Golay filter determines to be approximately 20 $m\,s^{-2}$ as the velocity increases to over 400 $km\,s^{-1}$ before the data gaps cause a large scatter in the derived kinematics (e.g., an artificial acceleration peak of >100 $m\,s^{-2}$). The initial increasing velocity profile up to a maximum in the range ~600 – 800 $km\,s^{-1}$ by ~20:00 UT agrees with that of Byrne et al. (2009) as shown in their Fig. 7. The quadratic (and linear) fits in CORIMP are not as prone to the scattering effects of the data gap, and thus derive a slightly lower maximum velocity range of ~500 – 550 $km\,s^{-1}$ and an acceleration in the range of approximately ±15 $m\,s^{-2}$. CACTus determined a linear velocity of 463 $km\,s^{-1}$ (in the range 227 – 744 $km\,s^{-1}$). SEEDS determined a linear velocity of 338 $km\,s^{-1}$ and an acceleration of 17.7 $m\,s^{-2}$ (in the C2 field-of-view). And CDAW determined a

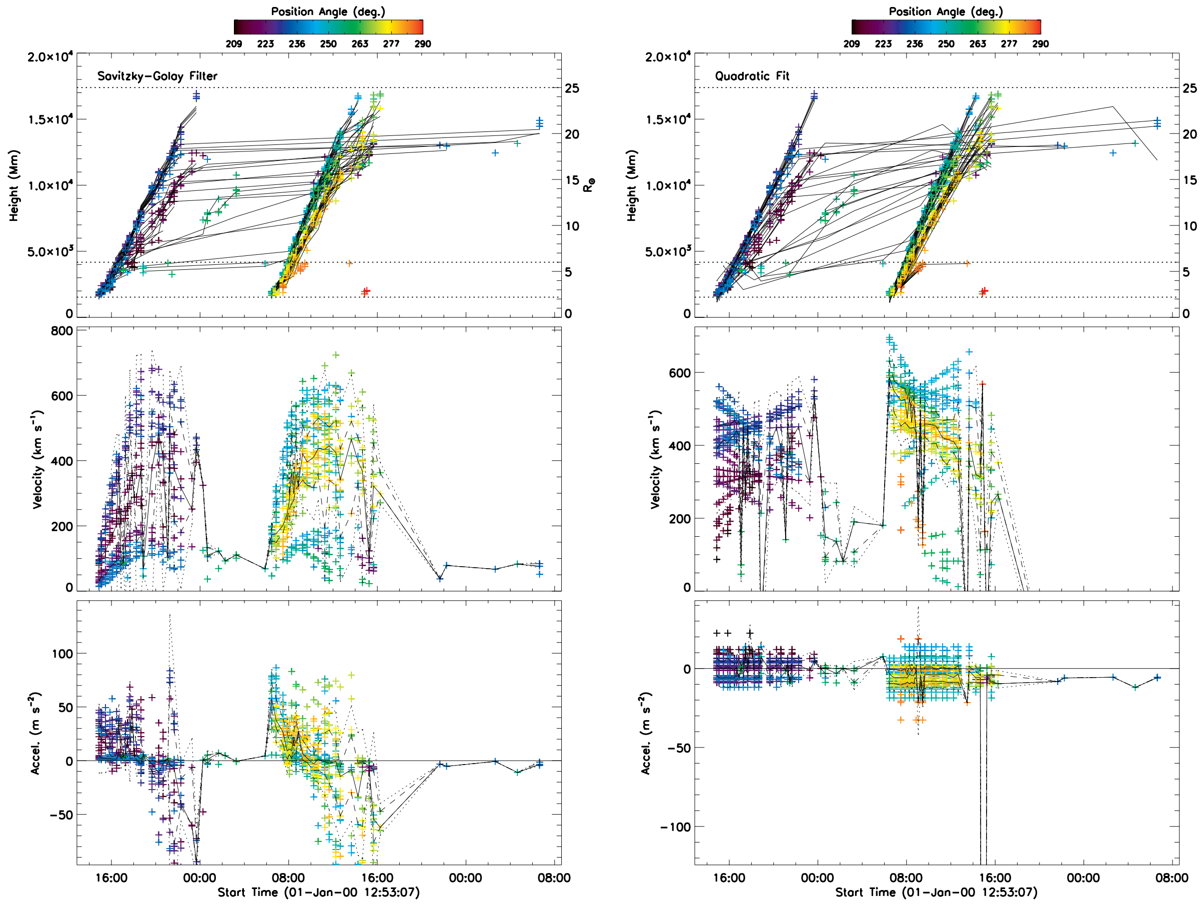

Fig. 4. Kinematic plots of the 2000 Jan. 02 CME from the automatic detection and tracking in the CORIMP catalog. The top plots show the height-time measurements with a colorbar to indicate the angular span of the data points. The middle and bottom plots show the velocity and acceleration profiles of the CME with the median (solid line), interquartile range (inner dashed lines) and upper and lower fences (outer dashed lines) over-plotted. The left plots are determined by a Savitzky-Golar filter applied to the height-time measurements with a 7-point moving window, while the right plots are determined with a second-order quadratic fit.

linear velocity of 668 $km\,s^{-1}$ and an overall acceleration of 23.1 $m\,s^{-2}$. These catalog measurements are listed in Table 2. Therefore, by inspection, the results of the CORIMP CME catalog are in relative agreement with the corresponding results of the other catalogs and manual analysis, and CORIMP is deemed robust and reliable.

3.3. Impulsive CME: 2000 April 23

The large and fast CME that erupted off the west limb of the Sun on 2000 Apr. 23 from ~12:54 UT in LASCO underwent a hugely impulsive acceleration as it exploded into the corona. CORIMP identified the bulk of the CME through the LASCO field-of-view to ~20 $R_\odot$ after which the CME front became too faint. Strong streamer deflections occurred to the north and south flanks of the CME,

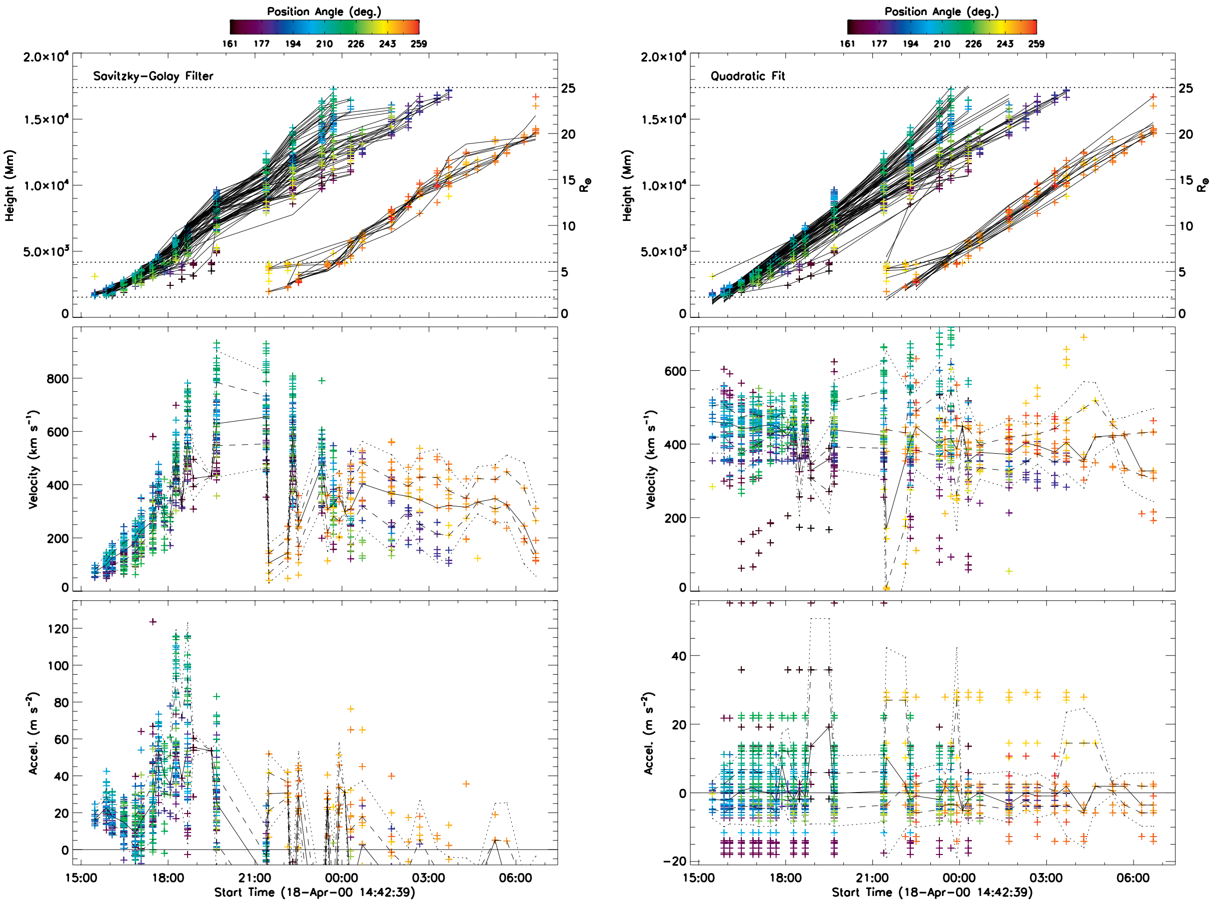

Fig. 5. Kinematic plots of the 2000 Apr. 18 CME from the automatic detection and tracking in the CORIMP catalog, as in Fig. 4.

Catalog	CPA [$deg.$]	AW [$deg.$]	Lin. Speed [$km\,s^{-1}$]	Accel. [$m\,s^{-2}$]
CORIMP	287	119^{125}	836^{1706}	-11^{50}_{-154}
CACTus	144	360	1114^{1849}_{245}	
SEEDS	275	130	594	-8.5
CDAW	281	360	1187	-48.5

Table 3. Catalog measurements of a CME observed by SOHO/LASCO on 2000 Apr. 23 from ∼12:54 UT

with very faint material visible as a full halo or shock around the east limb separate to the bulk flux-rope structure in the west. The CORIMP height-time measurements (Fig. 6) reveal an initial acceleration that the Savitzky-Golay filter determines to be $\gtrsim 150\,m\,s^{-2}$ dropping quickly to a range of approximately -100 to $0\,m\,s^{-2}$, as the velocity decreases from ~ 1000 to $500\,km\,s^{-1}$; though this is an underestimate since the filter overly smoothes the relatively under-sampled height-time measurements. The quadratic fits in CORIMP better handle this data and derive an initial velocity range

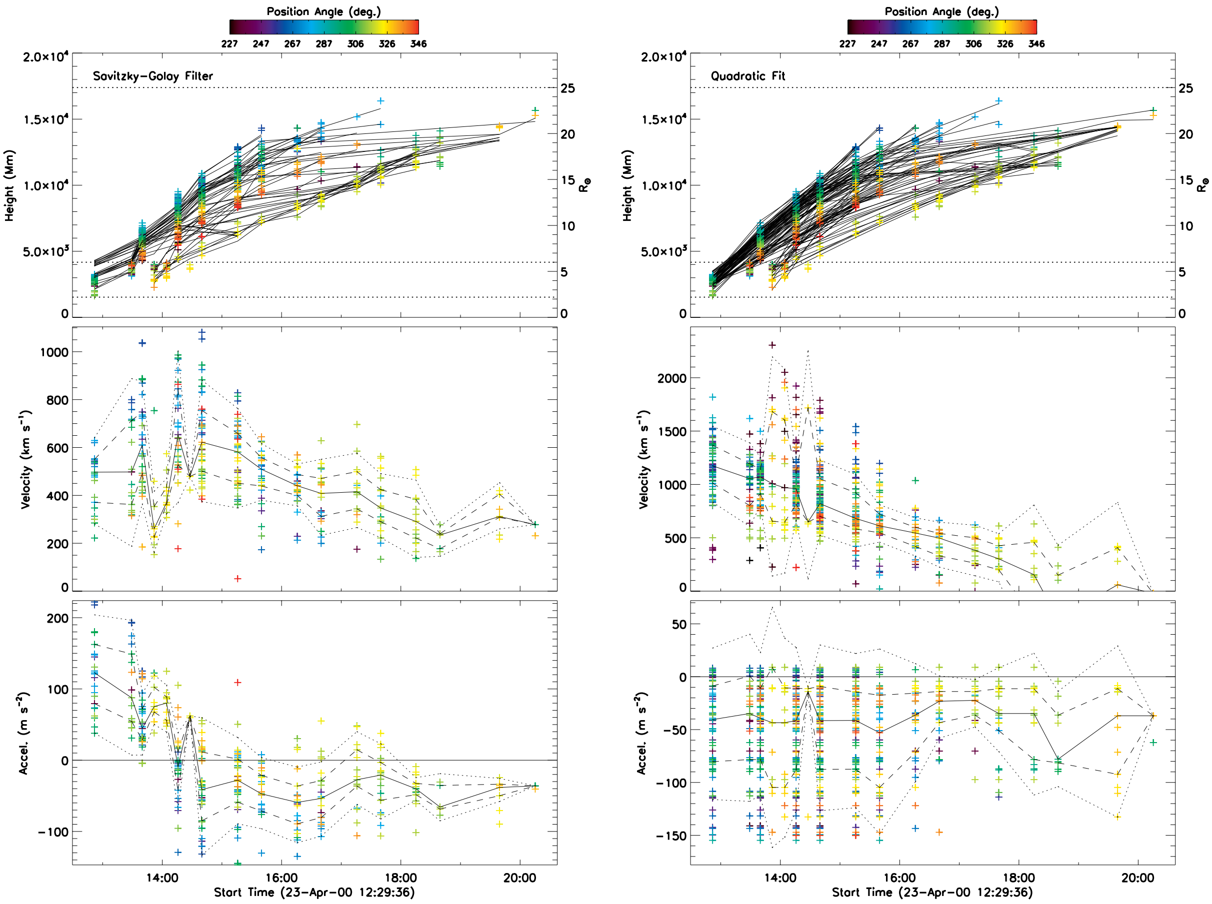

Fig. 6. Kinematic plots of the 2000 Apr. 23 CME from the automatic detection and tracking in the CORIMP catalog, as in Fig. 4.

of ~1200 – 1500 $km\ s^{-1}$, while the linear fits derive an initial velocity range of ~1000 – 1200 $km\ s^{-1}$, which are consistent with the measurements of Byrne et al. (2009) shown in their Fig. 8. The resulting deceleration is determined to have a median of approximately $-50\ m\ s^{-2}$, reaching as low as $-150\ m\ s^{-2}$. CACTus determined a linear velocity of 1114 $km\ s^{-1}$ (in the range 245 – 1849 $km\ s^{-1}$). SEEDS determined a linear velocity of 594 $km\ s^{-1}$ and a deceleration of $-8.5\ m\ s^{-2}$ (in the C2 field-of-view). And CDAW determined a linear velocity of 1187 $km\ s^{-1}$ and an overall deceleration of $-48.5\ m\ s^{-2}$. These catalog measurements are listed in Table 3. Therefore, by inspection and careful consideration of the low sampling of the event, the results of the CORIMP CME catalog are in relative agreement with the corresponding results of the other catalogs and manual analysis, and CORIMP is deemed robust and reliable.

3.4. Faint CME: 2001 April 23

The CME that erupted off the southwest limb of the Sun on 2001 Apr. 23 from ~12:54 UT in LASCO appeared relatively faint behind multiple streamers in the line-of-sight, some of which deflected especially along the southern flank of the CME. CORIMP identified the bulk of the CME

Catalog	CPA [*deg.*]	AW [*deg.*]	Lin. Speed [$km\,s^{-1}$]	Accel. [$m\,s^{-2}$]
CORIMP	232	72^{74}	187^{283}	3^{15}_{-13}
CACTus	231	88	459^{602}_{315}	
SEEDS	224	77	408	-46.6
CDAW	228	91	530	-0.7

Table 4. Catalog measurements of a CME observed by SOHO/LASCO on 2001 Apr. 23 from ~12:39 UT

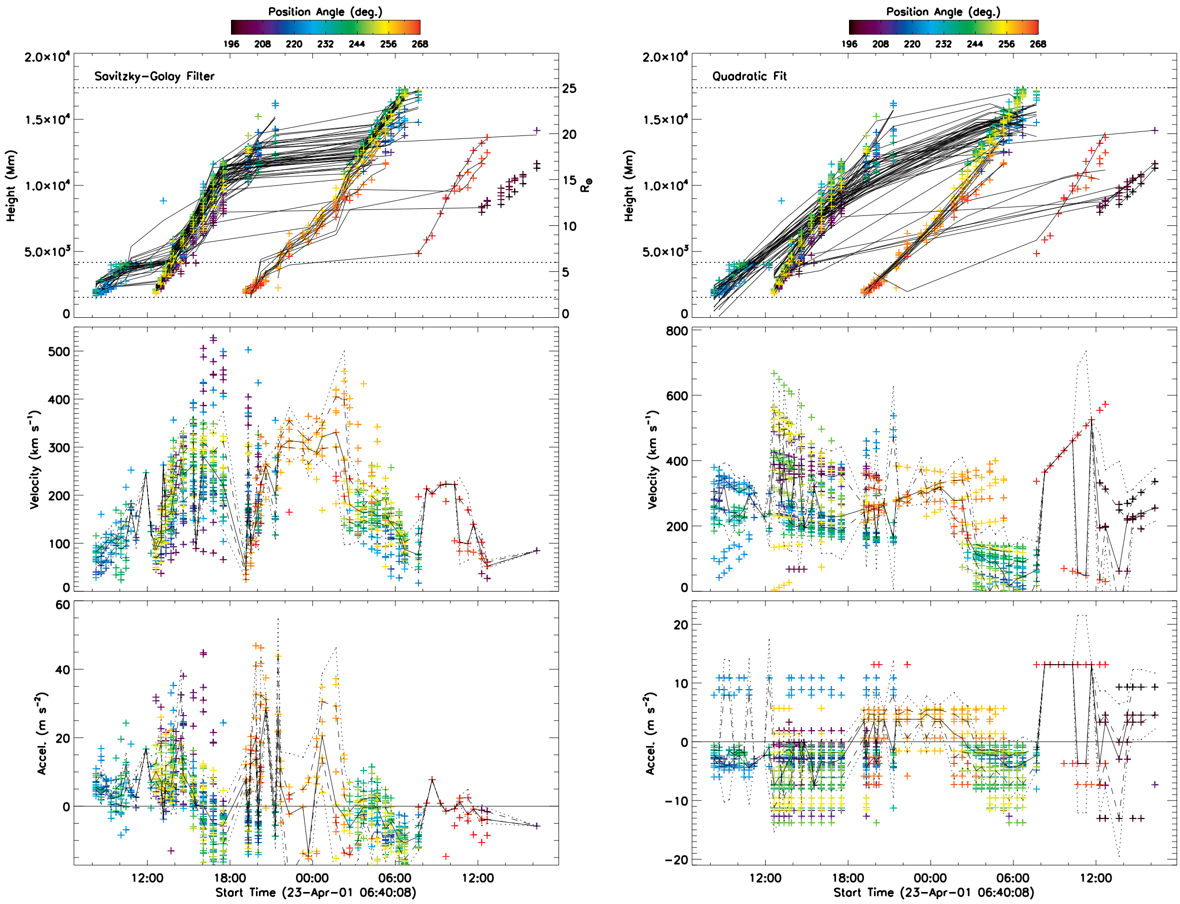

Fig. 7. Kinematic plots of the 2001 Apr. 23 CME from the automatic detection and tracking in the CORIMP catalog, as in Fig. 4.

through the LASCO field-of-view to ~20 $R_\odot$ after which the CME front became too faint. However, this CME was the first of two that occurred in close succession off the southwest limb, that CORIMP failed to separate due to their spatial and temporal overlap (plus some ejecta ahead of this CME was detected from ~08:16 UT). Therefore the kinematic profiles must be inspected before trusting the quoted catalog values listed in Table 4. Investigating the relevant portion of the plots in Fig. 7, in the time interval ~12:00 – 18:00 UT, the CORIMP height-time measurements reveal an initial acceleration that the Savitzky-Golay filter determines to be $\sim 10\, m\, s^{-2}$ dropping to scatter about zero as the maximum velocity levels off at $\sim 350\, km\, s^{-1}$; though this is an underestimate since the measurements are dominated by the material ahead of the CME front that the algorithm detected as part of the main event (from ~08:00 – 12:00 UT). The quadratic fits to the measurements are more dominated by the overall deceleration of the CME (approx. $-10\, m\, s^{-2}$ from 12:00 UT onwards) as the velocity drops from an initial range of $\sim 550 - 650\, km\, s^{-1}$ (consistent with Byrne et al. 2009 shown in their Fig. 9) to $\sim 400\, km\, s^{-1}$ by 18:00 UT; though this appears biased to lower values by the overlapping measurements of the second CME. The linear fits are less trustworthy as they tend to fit across the two CMEs and preceding ejected material, resulting in the underestimated CORIMP linear speed in Table 4. CACTus determined a linear velocity of $459\, km\, s^{-1}$ (in the range $315 - 602\, km\, s^{-1}$). SEEDS determined a linear velocity of $408\, km\, s^{-1}$ and a deceleration of $-46.6\, m\, s^{-2}$ (in the C2 field-of-view), however it failed to detect the CME front in the final frames which accounts for this erroneously large deceleration. And CDAW determined a linear velocity of $530\, km\, s^{-1}$ and an overall deceleration of $-0.7\, m\, s^{-2}$. These catalog measurements are listed in Table 4. Therefore, by inspection and careful consideration of the appropriate measurements of this event, the results of the CORIMP CME catalog are in relative agreement with the other catalogs and manual analysis where valid, and CORIMP is deemed robust albeit unreliable at separating overlapping events.

3.5. *Fast CME: 2002 April 21*

Catalog	CPA [$deg.$]	AW [$deg.$]	Lin. Speed [$km\, s^{-1}$]	Accel. [$m\, s^{-2}$]
CORIMP	235	154^{177}	1129^{2300}	61^{345}_{-619}
CACTus	322	352	1103^{1913}_{298}	
SEEDS	250	186	703	31.8
CDAW	282	360	2393	-1.4

Table 5. Catalog measurements of a CME observed by SOHO/LASCO on 2002 Apr. 21 from ~01:26 UT

The CME that erupted off the west limb of the Sun on 2002 Apr. 21 from ~01:27 UT in LASCO propagated very fast through the field-of-view. CORIMP identified the bulk of the CME through the LASCO field-of-view to $\sim 17\, R_\odot$ after which the CME front became too faint, and only the southern flank material continued to be detected. Figure 8 shows the CORIMP height-time measurements, which the Savitzky-Golay filter struggles to fit appropriately due to the small window-size available at each position angle (as the filter requires a minimum of 7 data points). The quadratic fits to the data reveal a high initial acceleration of $\gtrsim 1000\, m\, s^{-2}$ followed by a deceleration in the range of approximately -500 to $0\, m\, s^{-2}$. The velocity shows an initial range of $\sim 2000 - 2500\, km\, s^{-1}$ possibly reaching $\sim 3000\, km\, s^{-1}$ before dropping to $\sim 1000\, km\, s^{-1}$, which is consistent with the measurements of Byrne et al. (2009) shown in their Fig. 10. The linear fits also reveal an initial velocity range of $\sim 2000 - 2500\, km\, s^{-1}$ dropping to $\sim 1000\, km\, s^{-1}$. CACTus determined a linear velocity of $1103\, km\, s^{-1}$ (in the range $298 - 1913\, km\, s^{-1}$). SEEDS determined a linear velocity of $703\, km\, s^{-1}$ and an acceleration of $31.8\, m\, s^{-2}$ (in the C2 field-of-view), however these cannot be trusted as the CME front is only visible in two C2 frames. And CDAW determined a linear velocity of $2393\, km\, s^{-1}$ and

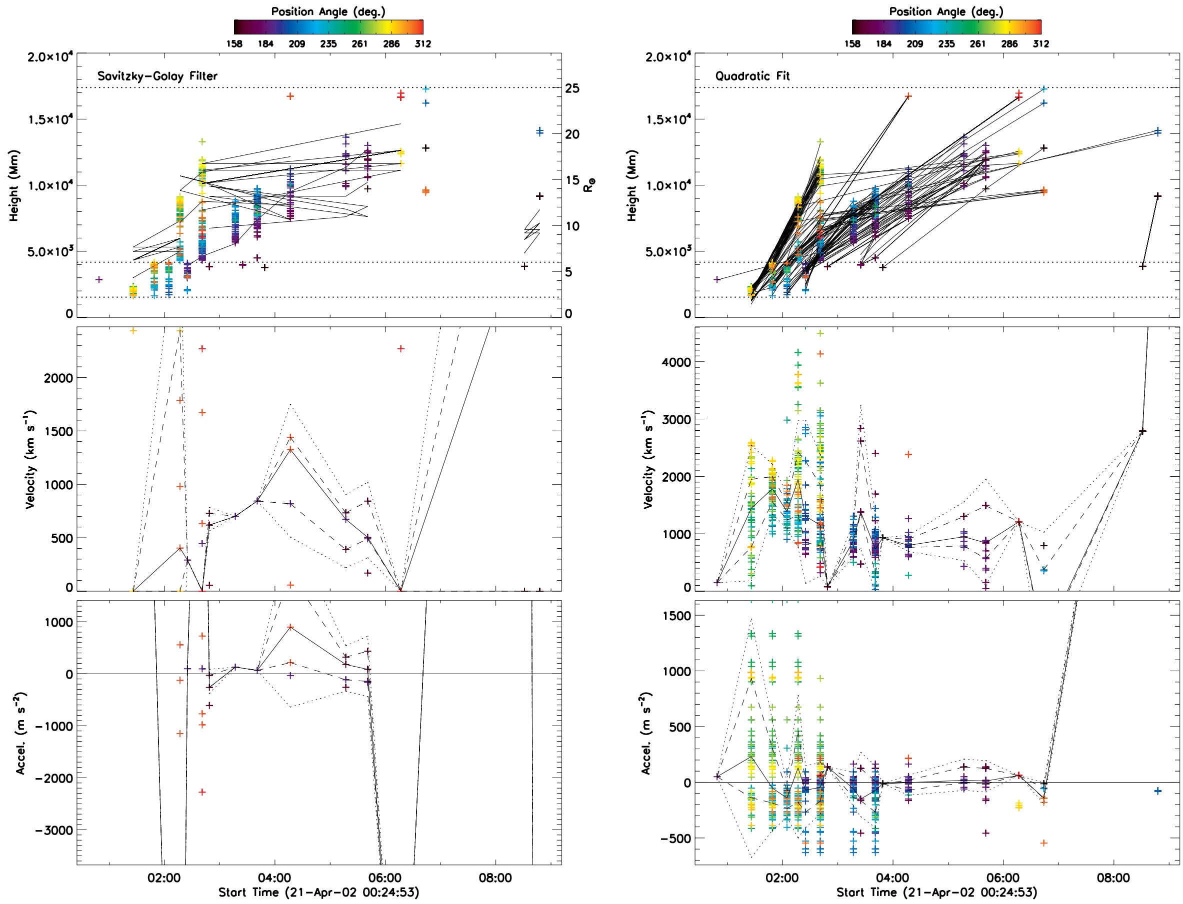

Fig. 8. Kinematic plots of the 2002 Apr. 21 CME from the automatic detection and tracking in the CORIMP catalog, as in Fig. 4.

an overall deceleration of $-1.4\,m\,s^{-2}$. These catalog measurements are listed in Table 5. Therefore, by inspection and careful consideration of the appropriate measurements of this event, the results of the CORIMP CME catalog are in relative agreement with the other catalogs and manual analysis where valid; and CORIMP is deemed reliable albeit problematic at handling low-sampled events (with the Savitzky-Golay filter anyway, as the quadratic and linear fits remain robust).

3.6. Flux-rope/Slow CME: 2004 April 1

The CME that erupted off the northeast limb of the Sun on 2004 Apr. 01 from ~23:05 UT in LASCO, exhibited a clear flux-rope structure and propagated relatively slowly. CORIMP identified the bulk of the CME through the LASCO field-of-view to $\sim 20\,R_\odot$ after which the CME front became too faint. Figure 9 shows the CORIMP height-time measurements, which are plentiful given the slow motion and clean detection of the event. These measurements reveal an initial acceleration that the Savitzky-Golay filter determines to be $\gtrsim 25\,m\,s^{-2}$ dropping to $0\,m\,s^{-2}$ by the time the CME reaches $\sim 15\,R_\odot$ and the maximum velocity levels off in the range $\sim 500-600\,km\,s^{-1}$. The quadratic fits to the data reveal a bulk velocity in the range $\sim 400-600\,km\,s^{-1}$, with an overall decelera-

Catalog	CPA [$deg.$]	AW [$deg.$]	Lin. Speed [$km\,s^{-1}$]	Accel. [$m\,s^{-2}$]
CORIMP	58	42^{44}	401^{502}	2^{18}_{-22}
CACTus	60	70	485^{829}_{244}	
SEEDS	60	59	261	19.7
CDAW	59	79	460	7.1

Table 6. Catalog measurements of a CME observed by SOHO/LASCO on 2004 Apr. 01 from ~23:04 UT

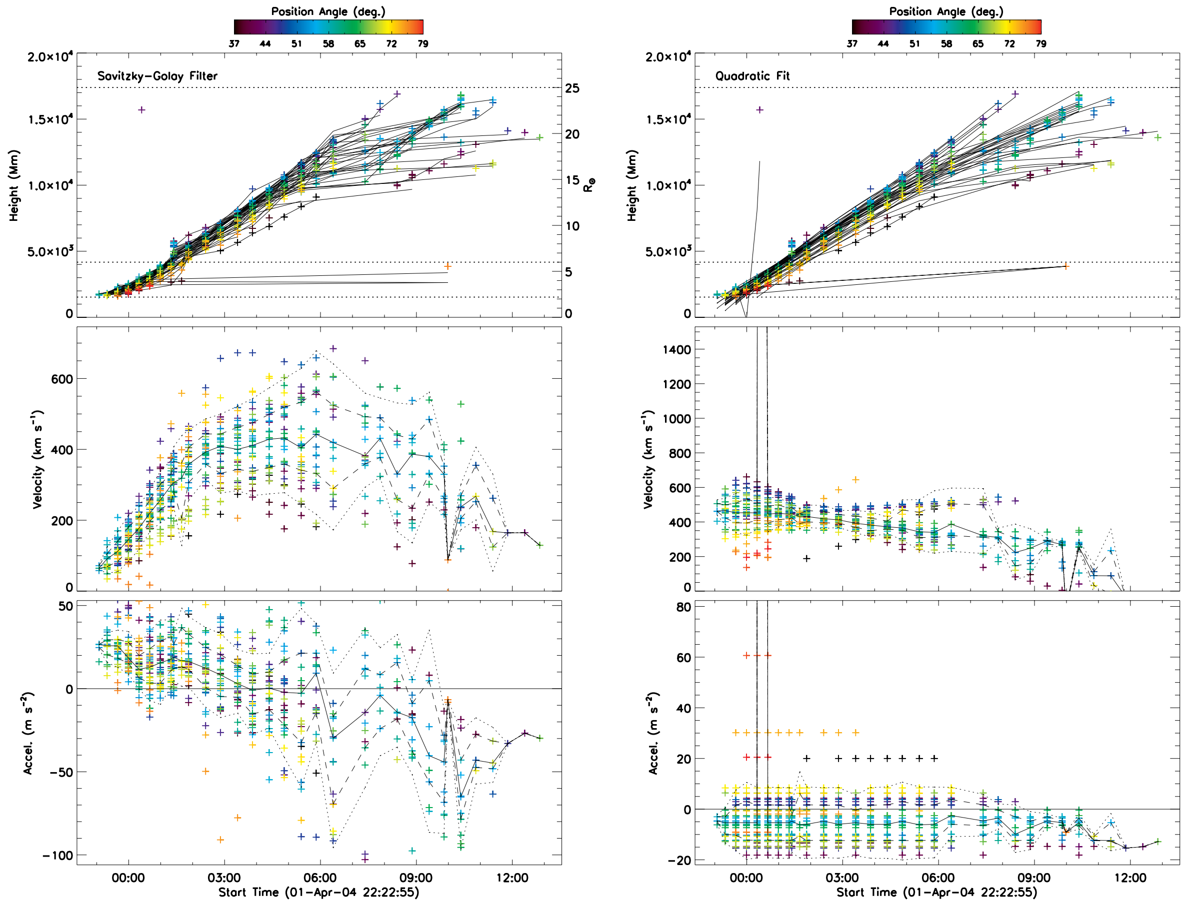

Fig. 9. Kinematic plots of the 2004 Apr. 01 CME from the automatic detection and tracking in the CORIMP catalog, as in Fig. 4.

tion of the CME of approximately $-5\,m\,s^{-2}$. The linear fits also produce a velocity in this range. These results are consistent with the measurements of Byrne et al. (2009) shown in their Fig. 11, though without reproducing the "staggered" velocity profile. CACTus determined a linear velocity of $485\,km\,s^{-1}$ (in the range $244-829\,km\,s^{-1}$). SEEDS determined a linear velocity of $261\,km\,s^{-1}$ and overall acceleration of $19.7\,m\,s^{-2}$ (in the C2 field-of-view). And CDAW determined a linear velocity of $460\,km\,s^{-1}$ and an overall acceleration of $7.1\,m\,s^{-2}$. These catalog measurements are listed in Table 6. Therefore, by inspection, the results of the CORIMP CME catalog are in relative agreement with the other catalogs and manual analysis, and CORIMP is deemed robust and reliable.

4. Separating Multiple CME Detections via *K*-means Clustering

The case studies in the previous section highlight a key issue in the automatic detection, tracking and cataloging of CMEs: namely the difficulties in distinguishing between multiple events that occur close together in space and time. The events of Sect. 3.1 and 3.4 demonstrate how the CORIMP catalog can fail to separately characterise CMEs that a human user would label as distinct events (although even this can be a non-trivial task since projection effects can make it hard to determine if two CMEs are truly merging in space or simply overlapping on the plane-of-sky). The reason for this, is that the thresholds in place to identify the beginning of a new CME detection cannot readily determine the end of a previous CME detection whose trailing material overlaps the subsequent CME material on the plane-of-sky. Other observational factors can help a human user distinguish the two, such as the differing speeds, densities, brightness and cohesiveness of the structure. However, these differences are too subtle to employ in an automated algorithm that must be able to characterise all manner of CMEs with a large variety of such properties; and so the overlapping CMEs are classified as a single event. While it is still possible to investigate their separate kinematic trends, as in Sect. 3.1 and 3.4, this is not ideal for accurately counting CMEs nor for reliably producing independent CME detection alerts. It therefore remains a challenge to employ some form of machine intelligence in cataloging CMEs, which can use the CME detection parameters to reveal instances when multiple CMEs occur together.

An initial effort to do this has been made using a clustering algorithm in the field of unsupervised machine learning. Specifically, the method of *k*-means clustering was investigated, which works by partitioning *n* observations into *k* clusters that are distinguished by minimising the within-cluster sum of squares, i.e., using Euclidean distance as a metric on the parameter space. It is well suited to generating globular, non-hierarchical, non-overlapping clusters, and may be computationally fast if *k* remains small. This approach could work with the parameters available in the CME detection analysis, such as the time, location/direction, size and speed of a CME. For example, the bulk of the overlapping CMEs may have different average propagation times as one may be proceeding slightly later in time than another. Or if the CMEs are propagating at different speeds, then a linear fit to each of their height-time profiles would have a different slope. Choosing these parameters of “average time” and “slope of a linear fit”, determined at every position angle in the span of the event detection, it is possible to cluster the height-time measurements into separate CMEs.

This is demonstrated in Figs. 10 and 11 for the two events discussed in Sects. 3.1 and 3.4, respectively. These figures show top plots of the *k*-means clustering algorithm applied to the parameters of “average time” and “slope of a linear fit”, where the means are plotted as black asterisks and the associated groups of points in separate clusters are plotted with different symbols and colours. The bottoms plots of these figures show the corresponding height-time profiles that have been separated according to their clusters. In Fig. 10 the results are shown for both $k = 3$ (left plots) and $k = 4$ (right plots), to illustrate the effect of changing k. By inspection, the clustering algorithm works well at distinguishing the separate events. The bulk measurements of the case-study CME beginning at ~06:06 UT on 2000 Jan. 02 are quite well clustered, though some of its later C3 measurements are wrongly determined as part of a separate CME. For this event the $k = 3$ case fares better at grouping

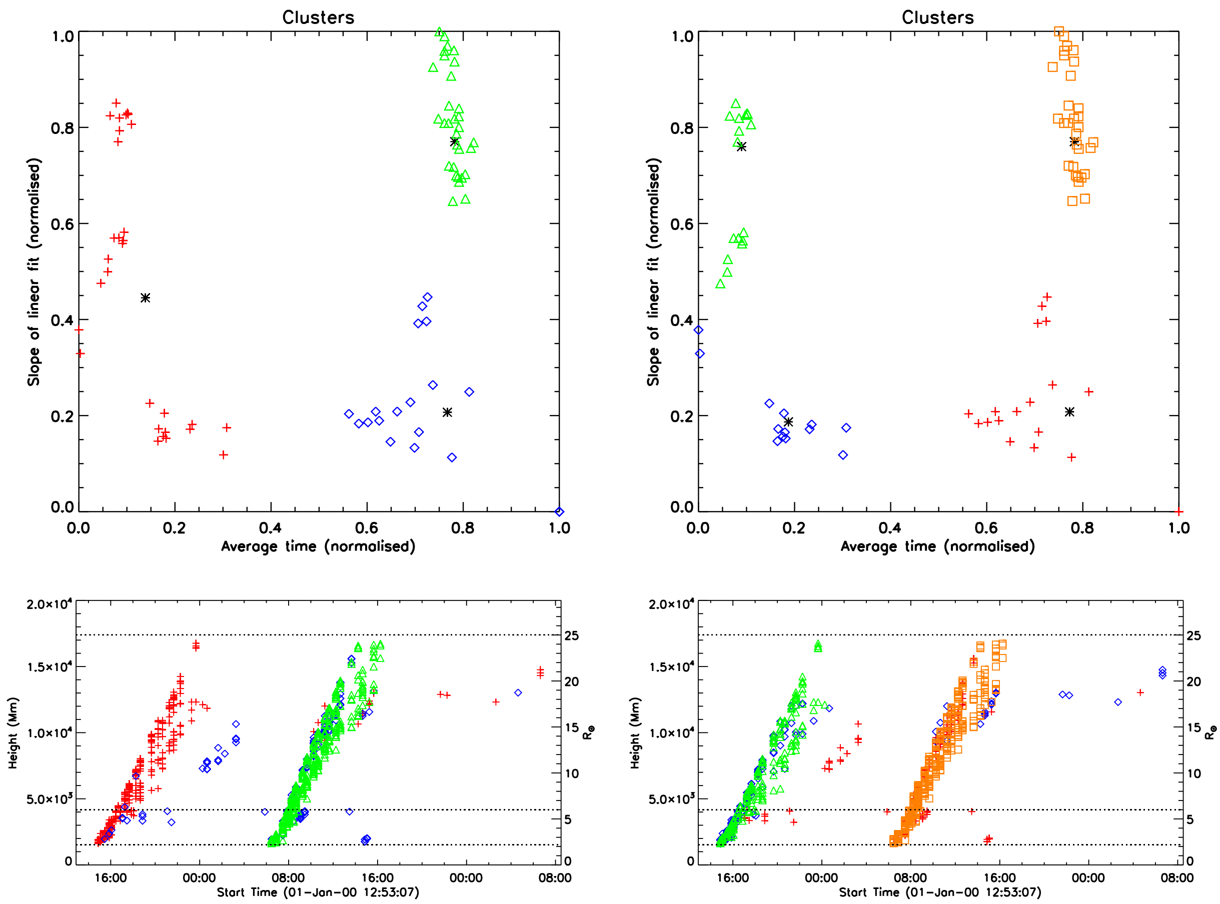

Fig. 10. A k-means clustering algorithm applied to the height-time data in Fig. 4 of the 2000 Jan. 01 CME, in an effort to distinguish the multiple CMEs detected as a single event due to their close proximity in space and time. The top plots show the resulting clusters for the instances of $k = 3$ (left) and $k = 4$ (right). The clustering is applied to the normalised parameters of the slope of a linear fit to the height-time profile at each position angle, and the mean time in the height-time profile at each position angle. The clusters are distinguished by different plot symbols and colours, and the black asterisks in the top plots indicate the mean of each cluster.

the CMEs, while the $k = 4$ case splits apart the profile of the first CME (shown as the green and blue datapoints in the bottom right plot of Fig. 10). Similarly in Fig. 11 the clustering algorithms go some way towards distinguishing multiple CMEs in the event detection on 2001 Apr. 23, but there are datapoints that are wrongly classified, even for the two instances of $k = 4$ shown for this event. These results highlight the difficulty in applying an automatic extraction of separate CME height-time profiles when detected so close together in space and time. Furthermore, there is an inherent limitation to k-means clustering by having to specify the number of clusters required from the data, which is not known a priori - especially not for an automated methodology such as in the CORIMP catalog. Further investigation into the parameters to be clustered, alternative clustering, or different machine learning algorithms, may produce better results.

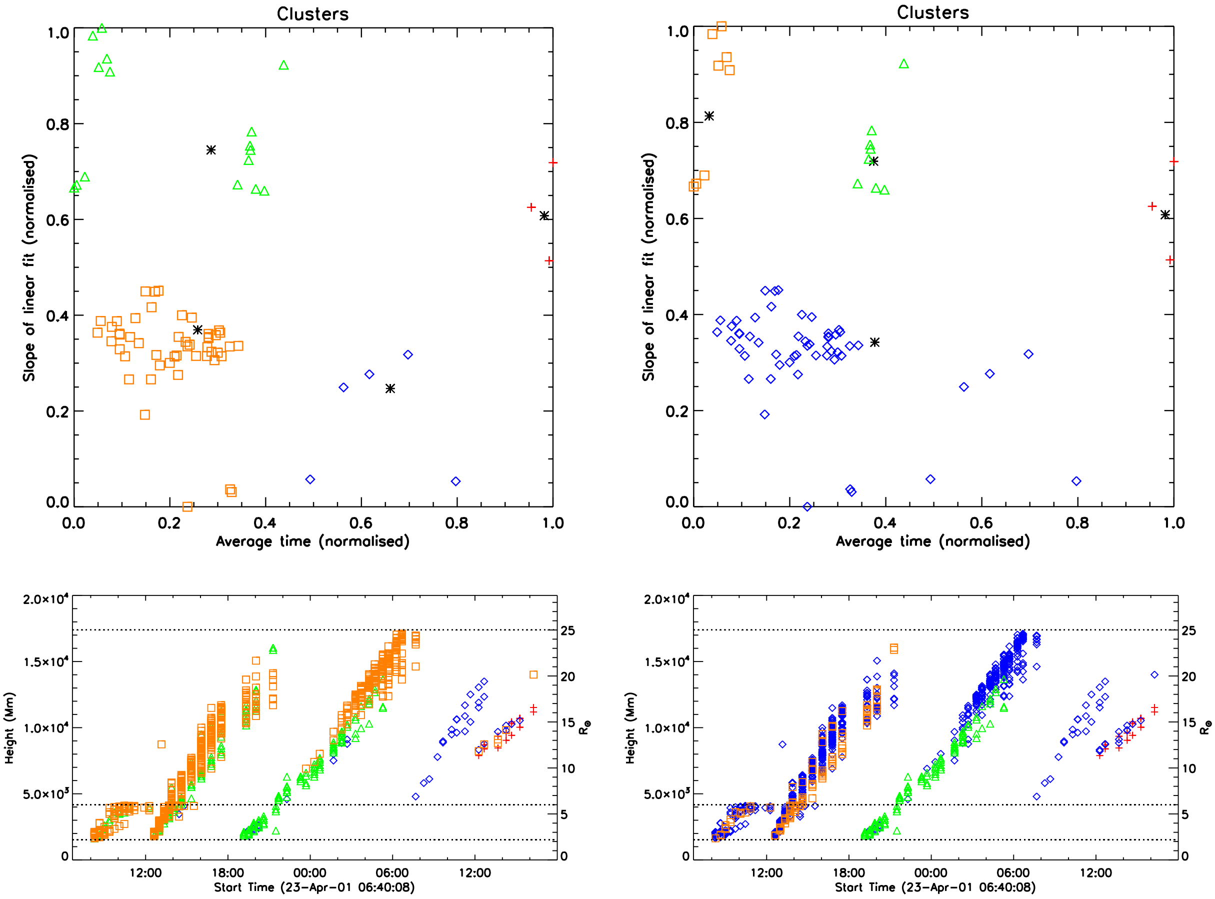

Fig. 11. A k-means clustering algorithm applied to the height-time data in Fig. 7 of the 2001 Apr. 23 CME, as in Fig. 10. The top plots show the resulting clusters for two instances of $k = 4$, to highlight the different possible results for the same k value, depending on how the algorithm converges on the optimum k-means. The bottom plots show the resulting height-time profiles corresponding to the separate CMEs distinguished by the clustering algorithm, shown by the different plot symbols and colours.

5. Conclusions

As the wealth of coronagraph data and CME observations has increased dramatically since the launch of SOHO in 1995, it has become important to develop robust and reliable methods of detecting and tracking CMEs in white-light images. Since CMEs are faint and transient phenomena that prove difficult to consistently isolate from the background corona, manual inspection of the images is open to interpretation and prone to user-specific biases. Similarly, it is challenging to fix the criteria and thresholds necessary in a computerised methodology for automating this task, although advances have been made to achieve this and provide the benefit of having a self-consistent catalog of results. Efforts to both manually and automatically catalog CMEs have been discussed in Sect. 2 with the aim of comparing how each fares in light of the newly developed CORIMP catalog, which was built to overcome some of the drawbacks of current catalogs. To this end, a selection of

CMEs was chosen from a previous study by Byrne et al. (2009), and the new results in the CORIMP catalog were investigated alongside the results of the automated CACTus and SEEDS catalogs and the manual CDAW catalog.

In the previous study of Byrne et al. (2009), the CMEs were characterised with the use of a multiscale edge-detection filter, whereby an ellipse was fitted to the isolated CME front and its apex tracked to produce height-time measurements. Since this approach avoided differencing the images, it was possible to quantify single-image uncertainties for the resulting height-time measurements, to be used for gauging a confidence interval on the derived CME kinematics. However, Byrne et al. (2013) demonstrated that the often-used method of numerical differentiation using 3-point Lagrangian interpolation, and its associated error propagation, is not wholly reliable at deriving the true CME kinematics. This motivated the use of the Savitzky-Golay filter along with quadratic and linear fits to the height-time measurements in CORIMP, across the angular span of the CME such that the statistical spread in the kinematics of each event may better indicate the true underlying trends. It is therefore warranted to compare these new automatically-generated results with the outputs of the other catalogs.

The spread of measurements along the angular span of the CME proves more useful than choosing a single fixed apex of the CME, because it propagates as an impermanent, evolving structure that can undergo various rates of expansion across the plane-of-sky. The variety of events chosen here as a subset of the thousands in the LASCO data is enough to demonstrate this. Having the angular spread of kinematics also provides insight to the bulk motion of the CME as well as its flanks and front: with the angular extent indicating the flanks and the upper values on the velocities indicating the CME front (usually the fastest part of its structure). Therefore a greater amount of information is available on the overall CME motion.

The Savitzky-Golay filter provides an indication of the kinematic trends that a first or second-order fit cannot necessarily produce. Since this filter is applied in a moving-window on the data-points, it can be problematic in cases of events with low-sampling (as in Sect. 3.5), but otherwise performs very well at automatically quantifying the different phases of acceleration of a CME. Therefore the dynamics of the eruption may be better quantified and understood.

While the robustness of the CORIMP catalog is clear (in so far as it can demonstrably produce results that are accurate and consistent across the data), there is a reliability issue that arises in cases of multiple CMEs that overlap in space and time. The problem with such cases is that another CME can erupt in the same direction as a previous one, close enough in time that the two detections are merged, as though the second CME were part of the trailing material of the first (as shown in Fig. 3). The opposite problem to this is that harsher thresholds would split apart single CMEs into multiple events, especially large CMEs with substantial trailing material. Indeed such problems can affect all automated catalogs, such that CORIMP appears to suffer from the former issue, while CACTus and SEEDS suffer from the latter. A form of unsupervised machine learning was explored, by applying a *k*-means clustering algorithm to certain parameters in the CME detections, in an effort to distinguish overlapping events. While these first results shown in Sect. 4 are promising, they highlight the difficulty of the task, and warrant further investigation. So for now, this issue is only overcome by a manual inspection of the data, as highlighted in the events of Sect. 3.1 and 3.4. In conclusion, any catalog should not be quoted blindly, as the thresholds cannot always distinguish the exact eruption that a user would isolate by eye. However, knowing this, CORIMP still offers the most rigorous details on the kinematics and morphologies of CMEs in a catalog to date, from which a user can infer a wealth of information.

Acknowledgements. The SOHO/LASCO data used here are produced by a consortium of the Naval Research Laboratory (USA), Max-Planck-Institut fuer Aeronomie (Germany), Laboratoire d'Astronomie (France), and the University of Birmingham (UK). SOHO is a project of international cooperation between ESA and NASA. The CACTus CME catalog is generated and maintained by the SIDC at the Royal Observatory of Belgium. The SEEDS CME catalog has been supported by NASA Living With a Star Program and NASA Applied Information Systems Research Program. The CDAW Data Center CME catalog is generated and maintained by NASA and The Catholic University of America in cooperation with the Naval Research Laboratory.

References

Brueckner, G. E., R. A. Howard, M. J. Koomen, C. M. Korendyke, D. J. Michels, et al. The Large Angle Spectroscopic Coronagraph (LASCO). *Sol. Phys.*, **162**, 357–402, 1995. 10.1007/BF00733434. 1

Byrne, J. P., P. T. Gallagher, R. T. J. McAteer, and C. A. Young. The kinematics of coronal mass ejections using multiscale methods. *A&A*, **495**, 325–334, 2009. 10.1051/0004-6361:200809811, 0901.3392. 3, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 5

Byrne, J. P., D. M. Long, P. T. Gallagher, D. S. Bloomfield, S. A. Maloney, R. T. J. McAteer, H. Morgan, and S. R. Habbal. Improved methods for determining the kinematics of coronal mass ejections and coronal waves. *A&A*, **557**, A96, 2013. 10.1051/0004-6361/201321223, 1307.8155. 2.1, 3, 5

Byrne, J. P., H. Morgan, S. R. Habbal, and P. T. Gallagher. Automatic Detection and Tracking of Coronal Mass Ejections. II. Multiscale Filtering of Coronagraph Images. *ApJ*, **752**, 145, 2012. 10.1088/0004-637X/752/2/145. 1, 3.1

Chen, P. F. Coronal Mass Ejections: Models and Their Observational Basis. *Living Reviews in Solar Physics*, **8**, 1, 2011. 1

Colaninno, R. C., and A. Vourlidas. Analysis of the Velocity Field of CMEs Using Optical Flow Methods. *ApJ*, **652**, 1747–1754, 2006. 10.1086/507943. 1

Davis, C. J., J. A. Davies, M. Lockwood, A. P. Rouillard, C. J. Eyles, and R. A. Harrison. Stereoscopic imaging of an Earth-impacting solar coronal mass ejection: A major milestone for the STEREO mission. *Geophys. Res. Lett.*, **36**, 8102, 2009. 10.1029/2009GL038021. 1

Domingo, V., B. Fleck, and A. I. Poland. The SOHO Mission: an Overview. *Sol. Phys.*, **162**, 1–2, 1995. 10.1007/BF00733425. 1

Gallagher, P. T., C. A. Young, J. P. Byrne, and R. T. J. McAteer. Coronal mass ejection detection using wavelets, curvelets and ridgelets: Applications for space weather monitoring. *Advances in Space Research*, **47**, 2118–2126, 2011. 10.1016/j.asr.2010.03.028, 1012.1901. 1

Gopalswamy, N., S. Yashiro, G. Michalek, G. Stenborg, A. Vourlidas, S. Freeland, and R. Howard. The SOHO/LASCO CME Catalog. *Earth Moon and Planets*, **104**, 295–313, 2009. 10.1007/s11038-008-9282-7. 1, 2.4

Goussies, N. A., M. E. Mejail, J. Jacobo, and G. Stenborg. Detection and Tracking of Coronal Mass Ejections Based on Supervised Segmentation and Level Set. *Pattern Recogn. Lett.*, **31**(6), 496–501, 2010. 10.1016/j.patrec.2009.07.011, URL http://dx.doi.org/10.1016/j.patrec.2009.07.011. 1

Howard, R. A., J. D. Moses, A. Vourlidas, J. S. Newmark, D. G. Socker, et al. Sun Earth Connection Coronal and Heliospheric Investigation (SECCHI). *Space Science Reviews*, **136**, 67–115, 2008. 10.1007/s11214-008-9341-4. 1

Howard, T. A., and S. J. Tappin. Statistical survey of earthbound interplanetary shocks, associated coronal mass ejections and their space weather consequences. *A&A*, **440**, 373–383, 2005. 10.1051/0004-6361:20053109. 1

Hundhausen, A. J. Sizes and locations of coronal mass ejections - SMM observations from 1980 and 1984-1989. *J. Geophys. Res.*, **98**, 13,177, 1993. 10.1029/93JA00157. 1

Illing, R. M. E., and A. J. Hundhausen. Observation of a coronal transient from 1.2 to 6 solar radii. *J. Geophys. Res.*, **90**, 275–282, 1985. 10.1029/JA090iA01p00275. 1

Koomen, M. J., C. R. Detwiler, G. E. Brueckner, H. W. Cooper, and R. Tousey. White Light Coronagraph in OSO-7. *Applied Optics*, **14**, 743–751, 1975. URL http://www.opticsinfobase.org/abstract.cfm?URI=ao-14-3-743. 1

Lugaz, N., and P. Kintner. Effect of Solar Wind Drag on the Determination of the Properties of Coronal Mass Ejections from Heliospheric Images. *Sol. Phys.*, 47, 2012. 10.1007/s11207-012-9948-1, 1204.3813. 1

MacQueen, R. M., A. Csoeke-Poeckh, E. Hildner, L. House, R. Reynolds, A. Stanger, H. Tepoel, and W. Wagner. The High Altitude Observatory Coronagraph/Polarimeter on the Solar Maximum Mission. *Sol. Phys.*, **65**, 91–107, 1980. 10.1007/BF00151386. 1

Morgan, H., J. P. Byrne, and S. R. Habbal. Automatically Detecting and Tracking Coronal Mass Ejections. I. Separation of Dynamic and Quiescent Components in Coronagraph Images. *ApJ*, **752**, 144, 2012. 10.1088/0004-637X/752/2/144. 1

Olmedo, O., J. Zhang, H. Wechsler, A. Poland, and K. Borne. Automatic Detection and Tracking of Coronal Mass Ejections in Coronagraph Time Series. *Sol. Phys.*, **248**, 485–499, 2008. 10.1007/s11207-007-9104-5. 1

Plunkett, S. P., B. J. Thompson, O. C. St. Cyr, and R. A. Howard. Solar source regions of coronal mass ejections and their geomagnetic effects. *Journal of Atmospheric and Solar-Terrestrial Physics*, **63**, 389–402, 2001. 10.1016/S1364-6826(00)00166-8. 1

Pulkkinen, T. Space Weather: Terrestrial Perspective. *Living Reviews in Solar Physics*, **4**, 1, 2007. 1

Robbrecht, E., and D. Berghmans. Automated recognition of coronal mass ejections (CMEs) in near-real-time data. *A&A*, **425**, 1097–1106, 2004. 10.1051/0004-6361:20041302. 1, 2.2

Schwenn, R., A. dal Lago, E. Huttunen, and W. D. Gonzalez. The association of coronal mass ejections with their effects near the Earth. *Annales Geophysicae*, **23**, 1033–1059, 2005. 1

Sheeley, N. R., Jr., D. J. Michels, R. A. Howard, and M. J. Koomen. Initial observations with the SOLWIND coronagraph. *ApJ*, **237**, L99–L101, 1980. 10.1086/183243. 1

St. Cyr, O. C., S. P. Plunkett, D. J. Michels, S. E. Paswaters, M. J. Koomen, et al. Properties of coronal mass ejections: SOHO LASCO observations from January 1996 to June 1998. *J. Geophys. Res.*, **105**, 18,169–18,186, 2000. 10.1029/1999JA000381. 1

Stenborg, G., and P. J. Cobelli. A wavelet packets equalization technique to reveal the multiple spatial-scale nature of coronal structures. *A&A*, **398**, 1185–1193, 2003. 10.1051/0004-6361:20021687. 1

Webb, D. F., and T. A. Howard. Coronal Mass Ejections: Observations. *Living Reviews in Solar Physics*, **9**, 3, 2012. 1

Yashiro, S., N. Gopalswamy, G. Michalek, O. C. St. Cyr, S. P. Plunkett, N. B. Rich, and R. A. Howard. A catalog of white light coronal mass ejections observed by the SOHO spacecraft. *Journal of Geophysical Research (Space Physics)*, **109**, 7105, 2004. 10.1029/2003JA010282. 1