Solar Physics
DOI: 10.1007/•••••-•••-•••-••••-•

Geoeffectiveness of Coronal Mass Ejections in the SOHO era

M. Dumbović[1] · A. Devos[2] · B. Vršnak[1] · D. Sudar[1] · L. Rodriguez[2] · D. Ruždjak[1] · K. Leer[3] · S. Vennerstrom[3] · A. Veronig[4]

Abstract

The main objective of the study is to determine the probability distributions of the geomagnetic *Dst* index as a function of the coronal mass ejection (CME) and solar flare parameters for the purpose of establishing a probabilistic forecast tool for the geomagnetic storm intensity. Several CME and flare parameters as well as the effect of successive-CME occurrence in changing the probability for a certain range of *Dst* index values, were examined. The results confirm some of already known relationships between remotely-observed properties of solar eruptive events and geomagnetic storms, namely the importance of initial CME speed, apparent width, source position, and the associated solar flare class. In this paper we quantify these relationships in a form to be used for space weather forecasting in future. The results of the statistical study are employed to construct an empirical statistical model for predicting the probability of the geomagnetic storm intensity based on remote solar observations of CMEs and flares.

Keywords: coronal mass ejections, solar flares, geomagnetic storms

1. Introduction

Interplanetary coronal mass ejections (ICMEs) are the drivers of the most intense geomagnetic storms (e.g., Gosling *et al.*, 1990; Koskinen and Huttunen, 2006). They carry enhanced magnetic fields, usually at a speed higher than the background solar wind, both parameters being crucial for generating geomagnetic storms in a process that involves magnetic reconnection with the Earth's magnetosphere. A geomagnetic storm occurs if the topology of the magnetic field in the ICME is favorable for reconnection, *i.e.*, if there is a strong southward

[1] Hvar Observatory, Faculty of Geodesy, University of Zagreb, Zagreb, Croatia; email: mdumbovic@geof.hr
[2] Royal Observatory of Belgium, Brussels, Belgium
[3] Technical University of Denmark, Copenhagen, Denmark
[4] IGAM/Institute of Physics, University of Graz, Graz, Austria

component of the magnetic field (Dungey, 1961; Russell, McPherron, and Burton, 1974; Akasofu, 1981). The connection between *in situ* properties of ICMEs and geomagnetic storms has been investigated in numerous studies considering different geomagnetic indices (see, *e.g.*, Huttunen *et al.*, 2005; Srivastava and Venkatakrishnan, 2004; Zhang *et al.*, 2007; Richardson and Cane, 2011; Yermolaev *et al.*, 2012; Verbanac *et al.*, 2013, and references therein). Monitoring of the near-Earth solar wind parameters can give a quite reliable prediction of the potentially harmful events. However, the warnings precede the event only by an hour (for spacecraft located at the Lagrangian point L1), providing very limited "response time" (Richardson and Cane, 2011). Therefore, it would be more useful to predict geoeffectiveness of ICMEs based on the remotely-observed CME/flare parameters.

Numerous studies dealing with the properties of geoeffective CMEs have been carried out including several attempts to construct geomagnetic storm prediction-models based on the remotely-measured properties of CMEs (Srivastava, 2005; Srivastava, 2006; Valach *et al.*, 2009; Kim *et al.*, 2010; Uwamahoro, McKinnell, and Habarulema, 2012). The studies led to the conclusion that the geoeffectiveness of CMEs is related to the following solar properties of CMEs and the associated solar flares: CME initial speed (Srivastava and Venkatakrishnan, 2004; Gopalswamy, Yashiro, and Akiyama, 2007), apparent angular width (Zhang *et al.*, 2003; Srivastava and Venkatakrishnan, 2004; Zhang *et al.*, 2007), source region location (Zhang *et al.*, 2003; Srivastava and Venkatakrishnan, 2004; Gopalswamy, Yashiro, and Akiyama, 2007; Zhang *et al.*, 2007; Richardson and Cane, 2010), the intensity of the CME-related flare (Srivastava and Venkatakrishnan, 2004), occurrence of successive CMEs (Gopalswamy, Yashiro, and Akiyama, 2007; Zhang *et al.*, 2007). However, most of the above studies did not take account for the false and missing alarms, *i.e.*, CMEs apparently having favorable solar properties, which did not produce geomagnetic storms, and the geomagnetic storms produced by CMEs with apparently non-favorable solar properties, respectively (see, *e.g.*, Schwenn *et al.*, 2005; Rodriguez *et al.*, 2009), since they considered only storm-related CMEs.

The aim of this study is to quantitatively analyze CME/flare parameters and their relationship to the storm intensity *viz.* the Disturbance Storm Time (Dst) geomagnetic index, derived from changes of the horizontal component of the geomagnetic field (see, *e.g.*, Verbanac *et al.*, 2011 and references therein). To account for both, geomagnetic storms and false alarms, we include a large sample of the events on the Sun which can be associated with the geomagnetic activity at the Earth without considering the interplanetary component. It is based on the data related to CME take-off and therefore is suited for the near real-time advance forecasting and warning. Based on the statistical analysis, we develop a model to construct a probability distribution for geoeffectiveness of an observed CME. However, this model suffers from some limitations regarding the forecast and number of false alarms, as discussed in Section 5 and will be analyzed further in future.

2. Data and event selection

The CME data was taken from the SOHO LASCO CME Catalog (Yashiro *et al.*, 2004, http://cdaw.gsfc.nasa.gov/CME_list/). The solar flare data was taken from the NOAA X-ray solar flare list (ftp://ftp.ngdc.noaa.gov/STP/space-weather/solar-data/solar-features/solar-flares).
CMEs were associated with solar flares in the time period 10 January 1996 – 30 June 2011 (hereafter "the SOHO era") using an automated method based on temporal and spatial criteria as described in Vršnak, Sudar, and Ruždjak (2005). The temporal criterion is used to associate a CME with all the flares within the ± 1 hour period of the CME liftoff time, where liftoff time is derived by back-extrapolation of the CME height-time plot (available on the CDAW website) to the solar surface assuming a linear speed. The spatial criterion associates a CME with all flares that were located within the opening angle of a CME, where the CME opening angle is a projection of the CME apparent width on the solar disc, centred around the central position angle of the CME obtained from LASCO-catalog. Therefore, the spatial criterion could not be used for halo CMEs (due to their apparent width of 360 degrees) and solar flares for which the location was not reported. Starting with a total of 16824 CMEs and 25907 flares in the SOHO era (reported by LASCO-catalog and NOAA Xray solar flare list, respectively) we first applied a temporal criterion to associate CMEs and flares. Then, the spatial criterion was used for the applicable events, resulting in a sample of 1392 CMEs and 1617 associated flares, meaning that some CMEs were associated with more than one flare. For those cases, the associated flare of the strongest intensity was chosen, resulting in 1392 CME-flare pairs. All but 38 pairs had a source position identified on the visible side of the Sun, meaning that they were front sided events. The remaining 38 CMEs for which the source position were not available, are halo CMEs therefore the association with flares was taken from the HALO CME SOHO LASCO catalog (http://cdaw.gsfc.nasa.gov/CME_list/halo/halo.html).

For the present analysis, we selected a subsample of the events, consisting of CMEs with speeds larger than 400 $\mathrm{km\,s^{-1}}$. From all the CMEs, we selected 211 events in order to equally cover the range of velocities (from 400 $\mathrm{km\,s^{-1}}$ to the fastest CMEs, *i.e.*, $v > 1500$ $\mathrm{km\,s^{-1}}$). Equal sampling was used due to the fact that 78% of CMEs in the sample of 1392 CME–flare pairs have speed less than 800 $\mathrm{km\,s^{-1}}$ (53% of CMEs have speed less than 500 $\mathrm{km\,s^{-1}}$). Further, previous studies have shown that faster CMEs are more geoeffective (*e.g.* Gopalswamy, Yashiro, and Akiyama, 2007). Therefore, using a random sample would include only a small number of large geomagnetic storms in the sample, *i.e.* most interesting events. For this purpose all fast CMEs ($v > 1500$ $\mathrm{km\,s^{-1}}$) were taken, including a total of 53 events, whereas for CMEs with 400 $\mathrm{km\,s^{-1}} < v < 1500$ $\mathrm{km\,s^{-1}}$ approximately 30 CMEs were randomly selected per bin of $\Delta v = 200$ $\mathrm{km\,s^{-1}}$. It is to be noted that the cases when the slower CMEs are likely to be overtaken by faster ones were also taken into consideration, however, these CMEs launched in quick succession were not treated as individual events. CMEs with less than three height-time measurements were discarded, due to uncertainty of the speed estimate. This criterion was relaxed in the case of

very fast CMEs ($v > 1500$ km s^{-1}), where only two height-time measurements are not unusual (see SOHO LASCO CME Catalog, Yashiro *et al.*, 2004, http://cdaw.gsfc.nasa.gov/CME_list/).

Using plots available on the SOHO-LASCO catalog, which associate the CME height-time measurement and the *Dst* index, we link the *Dst* events with CME-flare pairs (see an example shown in Figure 1). An extrapolation to the distance of 214 solar radii (approximately the distance from the Sun to Earth) was performed using CME "height-time" to derive a proxy time of arrival to the Earth. A *Dst* event was then sought in a specific time window, chosen to account for possible errors in the SOHO LASCO catalog speed measurements, influence of the drag (acceleration or decceleration by solar wind; see, *e.g.*, Cargill, 2004; Vršnak *et al.*, 2004; Vršnak *et al.*, 2013) and geometrical effects (ICME hitting with a flank; see, *e.g.*, Möstl and Davies, 2013). For CMEs in the speed range $v = 400 - 600$ km s^{-1} the time window starts 24 hours before and ends 36 hours after the proxy of the arrival time. For CMEs with speed $v > 600$ km s^{-1} the time window starts six hours before and ends 48 hours after the proxy of the arrival time. In this case a longer time beyond the time of estimated arrival was assumed because of the drag-decceleration effect and possible delayed impact of the flank, both of which depend on the speed of the CME (see *e.g.* Vršnak *et al.*, 2013 and references therein for the drag-decceleration effect and Möstl and Davies, 2013 for the flank-delayed impact). Within the time window, the *Dst* index was measured at the point where it reaches the minimum value (*Dst* timing). If there was no geomagnetic storm within the time window corresponding to a specific CME, any recognizable variation in the *Dst* index ($|Dst| \geq 10$ nT) closest to the proxy of arrival time (within the time window) was taken as the associated *Dst* level. The *Dst* timing in those cases is not a reliable parameter, therefore, the temporal aspect of geomagnetic storms (*e.g.* duration) is not included in the analysis. If there was no variation in *Dst* index throughout the time window which could be associated to a specific CME, the value of the *Dst* index at the proxy of arrival time was taken as the associated *Dst* level. It should be noted that there are no multiple associations between CME/flare pairs and *Dst* index values.

For each CME in the subsample of 211 events selected for study in the present paper a level of interaction with other CMEs was determined based on the following criteria:

- *the kinematical criterion* – interacting CMEs are associated with flares originating from the visible side of the Sun and their extrapolated kinematical curves cross or meet each other;
- *the timing criterion* – the liftoff of interacting CMEs is within a reasonable time window (≈ 2 days);
- *the source position/width criterion* – interacting CMEs originating from the same or neighbouring source region, *i.e.*, have close locations (unless halo and partial halo CMEs are involved, in which case this criterion was relaxed due to the fact that they have similar directions, *i.e.* they are presumably Earth-directed).

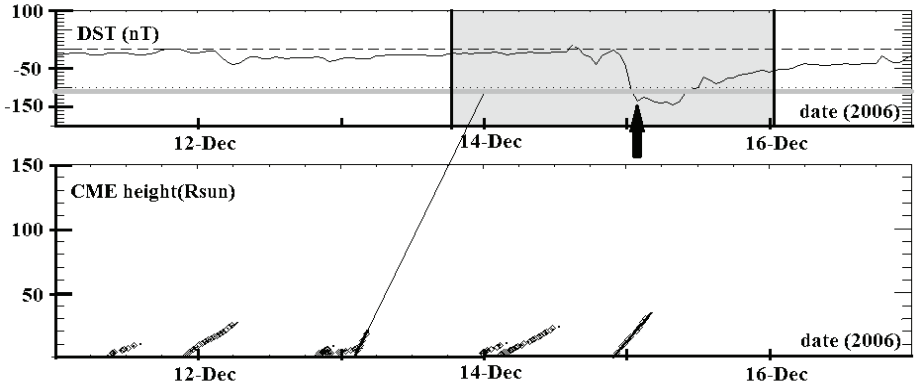

Figure 1. Association of a flare-related CME (first LASCO-C2 appearance, 13 December 2006, 02:54 UT) with a *Dst* event at Earth. The height-time curve (black solid line) is extrapolated to 1 AU (gray solid line). The shaded area represents the time window in which a *Dst* event was sought (six hours before and 48 hours after the proxy of arrival at Earth). Black arrow denotes the time at which the *Dst* level is measured.

We note that the listed criteria do not mean that CMEs necessarily interacted, they are used only to characterise the CMEs which are likely to interact. The kinematical criterion is based on the linear extrapolation of observed kinematical curves, without considering the drag effect. Furthermore, for simplicity we consider only flare-associated CMEs, for which the source location on the visible side of the Sun is identified. The timing criterion is introduced to prevent the unrealistically long chains of possibly interacting CMEs (*e.g*, a "CME1" kinematically interacts with a "CME2" that was launched a day before, which interacted with a "CME3" that started a day prior to "CME2", *etc.*). Finally, a source position/width criterion resolves cases where, *e.g*, two narrow CMEs from opposite limbs satisfy both kinematical and timing criterion, although they are unlikely to interact due to their different propagation directions. These criteria in many cases do not clearly indicate a possible interaction therefore we introduce the "interaction parameter" by which we specify four levels of "interaction probability":

- "SINGLE" (S) events - no interaction;
- "SINGLE?" (S?) - interaction not likely;
- "TRAIN?" (T?) - probable interaction;
- "TRAIN" (T) - interaction highly probable.

The determination of the interaction parameter is illustrated in Figure 2. The fastest CME (CME1, first appearance in LASCO-C2 15 June 2000, 07:54 UT) was a partial halo CME launched from N16W55; its proxy arrival time is marked with a black dot. It is preceded by three slower flare-related CMEs launched from source positions (chronologically backwards) N23W90 (CME2), N22W74 (CME3), and N21W69 (CME4), within a period of ≈ 2 days prior to the liftoff of the CME1. The extrapolated kinematical curve of CME1 crosses those of CME2 and CME4, but not of CME3. On the other hand, the extrapolated kinematical curves of CME3 and CME4 cross each other, whereas kinematical

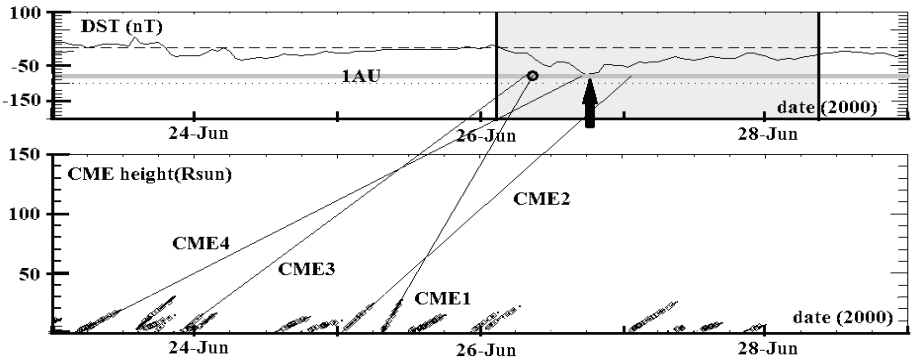

Figure 2. Association of a group of flare-related CMEs with a *Dst* event at Earth. We associate the fastest of the CMEs (CME1) with an interaction parameter "T?" (interaction likely) due to possible interaction with CMEs 2-4 based on the criteria described in Section 2 (for details see the main text). The *Dst* level is estimated as in Figure 1.

criterion for CME4 and CME2 is not met. Furthermore, CME2 is a narrow CME with source position at the limb, so we associate CME1 with an interaction level "T?" (interaction likely). The interaction parameter is assigned to each CME in the subsample of 211 events. We note that the whole CME train is then treated as one event that is characterized by solar parameters (*e.g.*, speed, width, flare association, *etc.*) of the fastest CME within a train.

For each event, *in situ* signatures were associated with *Dst* events. For this purpose we used the ICME list from (Richardson and Cane, 2010) available at http://www.srl.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm, *in situ* data from *Advanced Composition Explorer* satellite (ACE; Stone *et al.*, 1998) *Magnetometer* (MAG; Smith *et al.*, 1998) and *Solar Wind Electron, Proton, and Alpha Monitor* (SWEPAM; McComas *et al.*, 1998) instruments (http://www.srl.caltech.edu/ACE/ASC/level2/lvl2DATA_MAG-SWEPAM.html), and *in situ* data from *Wind* satellite *Magnetic Field Investigation* (MFI; Lepping *et al.*, 1995) and *Solar Wind Experiment* (SWE; Ogilvie *et al.*, 1995) instruments (http://wind.gsfc.nasa.gov/mfi_swe_plot.php). In this way we also checked if some of the geomagnetic storms with $|Dst| > 100$ nT were caused by a corotating interaction region (CIR) (Zhang *et al.*, 2003; Richardson *et al.*, 2006). We note that in the following analysis, CMEs associated with $|Dst| < 100$ nT are considered as non-relevant events of low geoeffectiveness. They either missed the Earth or did not produce a major storm. We note that some of them are in fact associated with CIRs, but from the prediction point of view, it is only relevant that they did not produce a geomagnetic storm with $|Dst| > 100$ nT, which is considered as the threshold for relevant strong geomagnetic activity.

In our sample of 211 CME–flare pairs the majority of the events were associated with ICMEs (57%), whereas 41% of events could not be associated with clear ICME signatures, *i.e.*, they were either CIRs, complex ejecta, or there was no *in situ* event at all. For 2% of events *in situ* data were not available due to measurement gaps. Out of 41% of events that were not associated with clear

ICME signatures, only one had $|Dst| > 100$ nT, however we did not discard it because it does not have clear CIR signatures as well.

3. Statistical analysis method

The selected sample of 211 CMEs/flares and associated Dst index provides numerous CME and flare parameters, as well as corresponding Dst value. Following the results of previous studies (*e.g.*, Zhang *et al.*, 2003; Srivastava and Venkatakrishnan, 2004; Gopalswamy, Yashiro, and Akiyama, 2007; Zhang *et al.*, 2007; Richardson and Cane, 2010; Richardson and Cane, 2011) we focus on specific parameters *viz.* the initial CME speeds and angular width, as well as solar flare soft X-ray class and location. In addition a level of interaction is also defined as a parameter since there are studies that indicate that interaction of CMEs can enhance their geoeffectivness (*e.g.*, Farrugia and Berdichevsky, 2004) and that most intensive storms are associated with trains of successive/multiple CMEs (Gopalswamy, Yashiro, and Akiyama, 2007; Zhang *et al.*, 2007; Möstl *et al.*, 2012; Mishra, Srivastava, and Chakrabarty, 2014). Distributions are used as a statistical tool for the analysis with the following bins: $|Dst| < 100\,\mathrm{nT}$, $100\,\mathrm{nT} < |Dst| < 200\,\mathrm{nT}$, $200\,\mathrm{nT} < |Dst| < 300\,\mathrm{nT}$, and $|Dst| > 300\,\mathrm{nT}$, where $|Dst|$ represents the magnitude of the Dst-index variation.

In order to check how $|Dst|$ distributions change for a specific key parameter, these key parameters were binned as well. For some key parameters the binning was obvious (*e.g.*, interaction parameter) as they are already discrete parameters. For continuous parameters all the bins have approximately the same number of events. The distribution mean, skewness, and kurtosis were calculated as relevant distribution parameters that depict the behavior of the $|Dst|$ distribution with the change in the (discrete) CME/flare parameter. The distribution skewness and kurtosis are coefficients derived from 3rd and 4th order moment of the distribution and represent asymmetry and peakedness/flatness coefficients, respectively.

The statistical significance of results was tested using two-sample t-test (2stt) at the 0.05 level (95% significance) for assuming dependence (equal variance assumed) and independence (equal variance not assumed) of the test samples. Due to the fact that 2stt is based on the normality assumption, *i.e.*, requires certain sample sizes, nonparametric significance tests were also performed, namely Kolmogorov-Smirnov and Mann-Whitney U-test, but there were no notable differences. In addition, no significant differences were noticed between the dependence and independence assumptions. Therefore we present only 2stt results for unequal variance (*i.e.* assumption of independence).

First, a general distribution of all measured values of the Dst index was performed, for two different types of measurements. Namely, in the $Dst-$time plot a geomagnetic storm is seen as a decrease in the Dst index, where the intensity of the storm is given by the magnitude of this depletion. The magnitude of this decrease was measured in two ways: the total magnitude ("total $|Dst|$"), measured from reference value 0, and relative magnitude ("relative" $|Dst|$), measured from the reference value at the start of the storm, *i.e.*, the

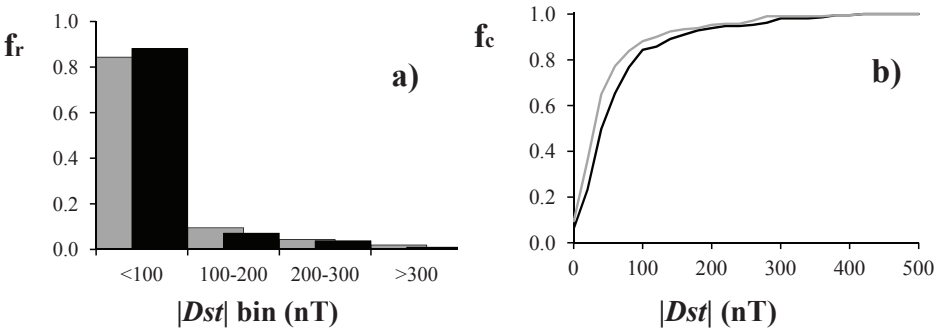

Figure 3. Distribution (a) and cumulative distribution (b) of the Dst-index variation (gray – total $|Dst|$; black – relative $|Dst|$).

amplitude of the Dst-index variation. We examine how the measurement process can affect the results, *i.e.*, is there a difference in considering the total or relative magnitudes. The statistical analysis shows that the distributions of total $|Dst|$ and relative $|Dst|$ are somewhat different, with relative $|Dst|$ shifted to lower values (Figure 3). The mean values are 68 nT and 53 nT, respectively, and are found to be significantly different at the 0.05 significance level with a 2stt. Total $|Dst|$ is usually larger than relative $|Dst|$, mostly due to the recovery of the preceding geomagnetic storms. While the total $|Dst|$ includes effects of the preceding storm, they are excluded in relative $|Dst|$, since it is measured from the onset point. Therefore, we focus our study on relative $|Dst|$, as a more realistic measure of storm strength. It should be noted though that the same analysis was repeated for total $|Dst|$ as well and similar results were obtained, therefore we do not present them here (the distributions were systematically shifted towards somewhat larger amplitudes, however, with the similar overall behavior).

In general, the $|Dst|$ distribution is highly asymmetric with over 80% of non geo-effective events ($|Dst| > 100$ nT for only 20 events). On the other hand, there are 110 HALO CMEs (52 %) and 140 CMEs with v >800 km s^{-1} (66 %), i.e. our sample contains a large number of false alarms. This is to be expected, because the sample was chosen based on the CME observations, where only a subset of CMEs caused geomagnetic storms. Therefore, although false alarms were not studied directly, their influence was taken into account (since they constitute a substantial part of the sample).

4. Results of the statistical analysis

In this section we analyze and discuss the relationship between CME/flare properties derived from remote solar observations and the $|Dst|$ levels at the Earth. In particular, we investigate the relationship between $|Dst|$ levels and the following solar parameters:

- *CME initial speed* (Section 4.1; Figures 4 and 5; Table 1);
- *CME/flare source position* (Section 4.2; Figures 6 and 9a; Table 2a);

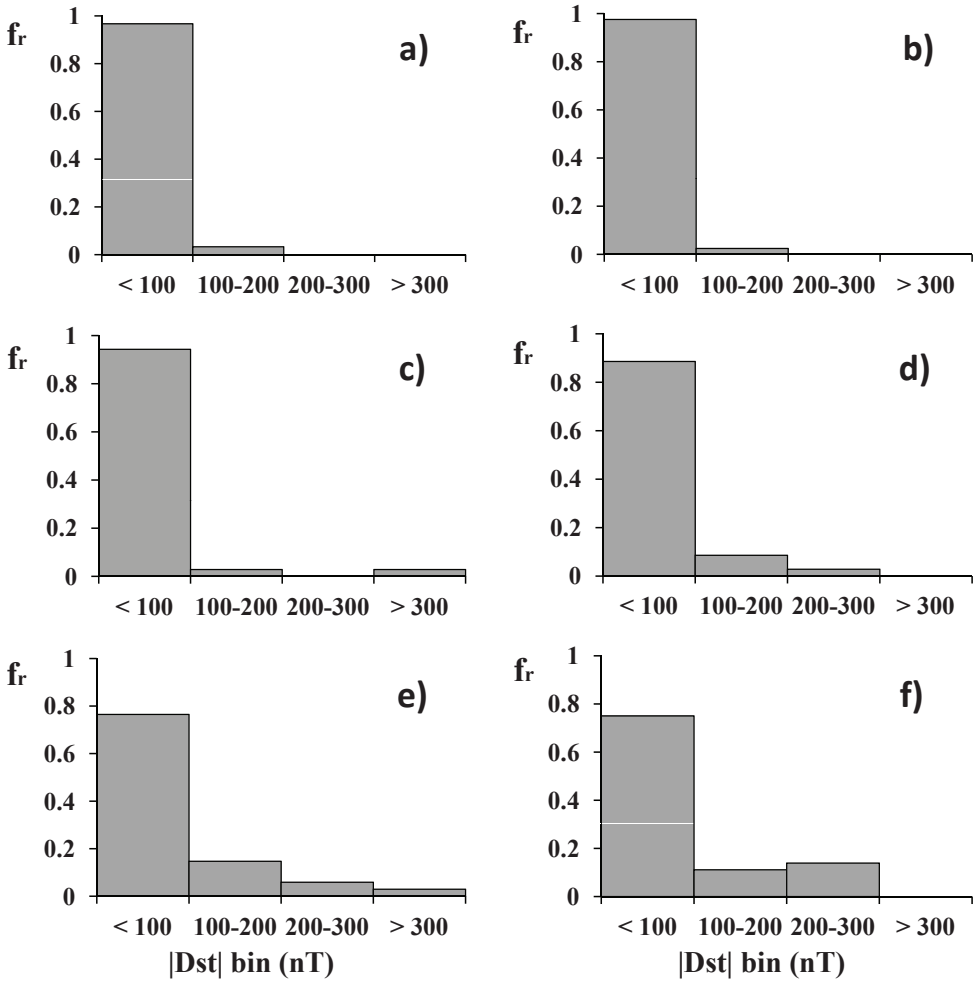

Figure 4. $|Dst|$ relative frequencies for different bins of CME speed, v: a) 400 – 600 km s^{-1}; b) 600 – 800 km s^{-1}; c) 800 – 1000 km s^{-1}; d) 1000 – 1200 km s^{-1}; e) 1200 – 1700 km s^{-1}; f)>1700 km s^{-1}.

- *CME–CME interaction parameter* (Section 4.3; Figures 7 and 9b; Table 2b);
- *CME angular width* (Section 4.4; Figures 8a and 9c; Table 3a);
- *solar flare class* (Section 4.5; Figures 8b and 9d; Table 3b).

In Section 4.6 we investigate the influence of combined solar parameters described in Sections 4.1 – 4.5 on the storm intensity ($|Dst|$ levels), whereas in Section 4.7 we investigate the overall behavior of the Dst index and compare it with the CME/flare activity to validate the sample and the statistical analysis.

Table 1. The two sample t-test significance levels of the difference between $|Dst|$ mean values in different bins of CME speed, v, with unequal variance assumed. Unless marked with an asterisk, the value states that the mean of the two samples are not significantly different; ** denotes that the significance of the difference is $> 95\%$; * denotes that the significance of the difference is $> 90\%$.

	v bins				
	bin1[1]	bin2	bin3	bin4	bin5
bin6	$2 \cdot 10^{-4}$**	$9 \cdot 10^{-4}$**	0.02**	0.01**	0.44
bin5	0.02**	0.05**	0.22	0.10*	—
bin4	0.30	0.73	0.66	—	—
bin3	0.16	0.42	—	—	—
bin2	0.33	—	—	—	—

[1]bins 1–6 represent different speed ranges in km s^{-1}: 400-600 (bin1), 600-800 (bin2), 800-1000 (bin3), 1000-1200 (bin4), 1200-1700 (bin5), and >1700 (bin6)

4.1. CME initial speeds

The first CME parameter analyzed is 1st order (linear) CME speed, v, derived from LASCO C2 and C3 images. Although this is plane-of-sky speed and subject to projection effects, it can be related to the radial speed of the CME (*e.g.*, Schwenn *et al.*, 2005; Gopalswamy *et al.*, 2012) and can therefore be taken as its proxy. The benefit is that this parameter can be relatively easily and quickly derived using L1 coronagraphic images. The events in our data sets were categorized into six different CME speed bins, with the following CME speed ranges: 400 – 600 km s^{-1} (bin 1), 600 – 800 km s^{-1} (bin 2), 800 – 1000 km s^{-1} (bin 3), 1000 – 1200 km s^{-1} (bin 4), 1200 – 1700 km s^{-1} (bin 5), and $v > 1700$ km s^{-1} (bin 6). The number of events in each bin is 36, 34, 35, 35, 41, and 30, respectively. In Figure 4 the $|Dst|$ distributions are presented for each CME speed bin, using the previously defined bin sizes. The mean value, skewness and kurtosis of the distributions were calculated to quantitatively examine changes in the $|Dst|$ distribution for different CME speed ranges (Figures 5a, 5c, and 5d, respectively). Furthermore, to test the differences between $|Dst|$ distributions (*i.e.*, whether they represent statistically different samples) a two sample t-test between each pair of $|Dst|$ distributions was applied. The results are presented in Table 1.

It can be seen in Figures 4a and 4b that for $v < 800$ km s^{-1} the distribution is restricted to $|Dst| < 200$ nT and is mostly contained within $|Dst| < 100$ nT. When compared with distributions in Figures 4c – 4f, these distributions lack the tail. This is also seen in the behavior of the distribution parameters (black dots in Figures 5a – d): the value of the mean, skewness and kurtosis first increase with increasing CME speed, up to the 800 – 1000 km s^{-1} bin. Then, as the speed range increases from the intermediate range speed bins (800 – 1000 km s^{-1}) to the higher range speed bins (> 1700 km s^{-1}) the distribution loses peakedness, as the values of skewness and kurtosis decrease, whereas the value of the distribution mean still increases. Changing the range of the CME speed leads to a change in the corresponding $|Dst|$ distribution in a way that from a

distribution the events are grouped in first two $|Dst|$ bins, the distribution first obtains the tail (mean, skewness and kurtosis increase), and then starts filling the tail (mean increases, whereas skewness and kurtosis decrease). This shift of distribution towards larger $|Dst|$ bins is also evident from the behavior of the distribution mean (Figure 5a), which can be approximated with a linear function (correlation coefficient, cc=0.96), although due to small number of data points this correlation should be taken with caution.

In addition, alternative speed bins were made, to substantiate our results. The alternative CME speed bins cover ranges: 400 – 700 $\mathrm{km\,s^{-1}}$ (52 events), 700 – 1000 $\mathrm{km\,s^{-1}}$ (54 events), 1000 – 1500 $\mathrm{km\,s^{-1}}$ (52 events), and $v > 1500$ $\mathrm{km\,s^{-1}}$ (53 events). The distribution parameters for these alternative distributions are shown as gray dots in Figures 5b – d. Including the distribution parameters of this alternative speed binning does not change the result notably, on the contrary, they follow the same trend.

Consequently, this means that faster CMEs have higher probabilities of producing strong geomagnetic storms, and furthermore, that slow CMEs ($v < 600$ $\mathrm{km\,s^{-1}}$) are not likely to produce intense storms ($|Dst| > 200$ nT) unless they are involved in a CME–CME interaction with a faster CME. The latter comes from the fact that interacting CMEs in the sample are related to the CME speed of the fastest CME in the train. It should be noted though that one parameter alone (e.g. CME initial speed) does not determine the geoeffectiveness of CMEs (as will be demonstrated in following sections). Therefore, distributions shown in Figure 4 are not a suitable measure of CME geoeffectiveness probability. The two sample t-test analysis reveals that there is no significant difference between two neighbouring speed bins (or several, as we go to lower speed bins) indicating that the $v - |Dst|$ relationship should be the most significant for very fast CMEs, whereas for slow CMEs it is unclear.

4.2. CME/flare source position

Several aspects of CME/flare source position were analyzed. First, the events were categorized by the quadrant in which the CME/flare source position is found (northeast, northwest, southeast, and southwest). It was confirmed by a two sample t-test that there is no difference in the samples from different quadrants. Next, it was investigated whether there is an asymmetry regarding the north/south and west/east source position of the CME/flare. Although a small difference in the $|Dst|$ distributions was observed between the west and east hemispheres, the two sample t-test could not confirm the differences.

Finally, a source distance from the solar disc centre, r, was investigated as a key parameter, ranging from 0 to 1 (in units of solar radii). Similarly as with CME speed, v, the events were optimally categorized into four bins: $r < 0.4$ (bin 1), $0.4 < r < 0.6$ (bin 2), $0.6 < r < 0.8$ (bin 3), and $r > 0.8$ (bin 4). The number of events in each bin is 45, 53, 53, and 60, respectively. For events involved in (possible) CME–CME interaction the source region of the fastest CME was taken as the relevant one. For each range of r, a $|Dst|$ distribution was made, using criteria of $|Dst|$ bins as discussed in Section 4.1. This resulted in four $|Dst|$ distributions for different r ranges of the CME/flare source region (Figure 6).

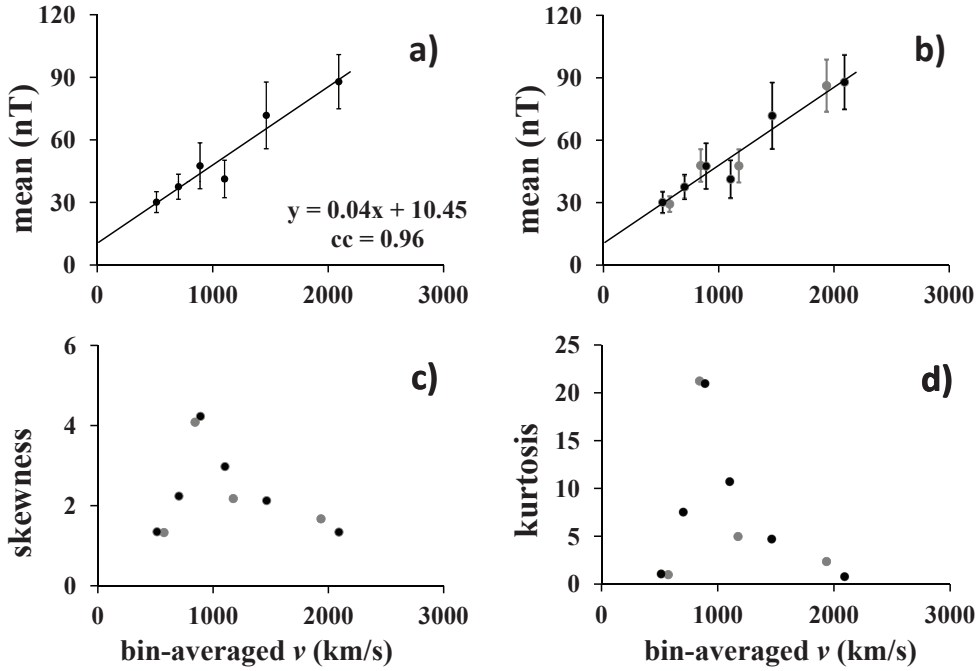

Figure 5. $|Dst|$ distribution parameters as a function of the bin-averaged value for the CME speed, v, in km s^{-1}. Black and gray dots mark the values for two different speed bins. While the black dots correspond to speed bins: 400 – 600 km s^{-1}, 600 – 800 km s^{-1}, 800 – 1000 km s^{-1}, 1000 – 1200 km s^{-1}, 1200 – 1700 km s^{-1}, and $v > 1700$ km s^{-1}, the gray dots correspond to speed bins: 400 – 700 km s^{-1}, 700 – 1000 km s^{-1}, 1000 – 1500 km s^{-1}, and $v > 1500$ km s^{-1}. Error bars in a) and b) represent confidence intervals, whereas the straight line shows a linear fit to data marked with black dots.

The results of the two sample t-test between each pair of $|Dst|$ distributions are presented in Table 2a.

It can be seen from Figure 6 that as the distance of the source region of the CME/flare from the disc centre increases the distribution loses the tail and for $0.6 < r < 0.8$ is restricted to $|Dst| < 200$ nT. This is somewhat expected and in agreement with numerous previous studies, where CMEs closer to the centre of the disc were found to be more geoeffective (*e.g.*, Zhang *et al.*, 2003; Srivastava and Venkatakrishnan, 2004; Gopalswamy, Yashiro, and Akiyama, 2007; Richardson and Cane, 2010). However, for the near-limb events the distribution again increases in the tail, showing that limb CMEs can also be highly geoeffective, as pointed out in, *e.g.*, Schwenn *et al.* (2005) and Cid *et al.* (2012). The two sample t-test shows significant differences only for the bin around the solar disc centre ($r < 0.4$), however, loss of significance for other bins does not seem stochastic, since there is a decrease in significance as we go towards the near-limb source locations (Table 2a). It can be seen in Figure 9a how the distribution mean decreases with the increasing distance from the disc centre, following approximately a power law (the curve is given illustratively in Figure 9a to guide the eye).

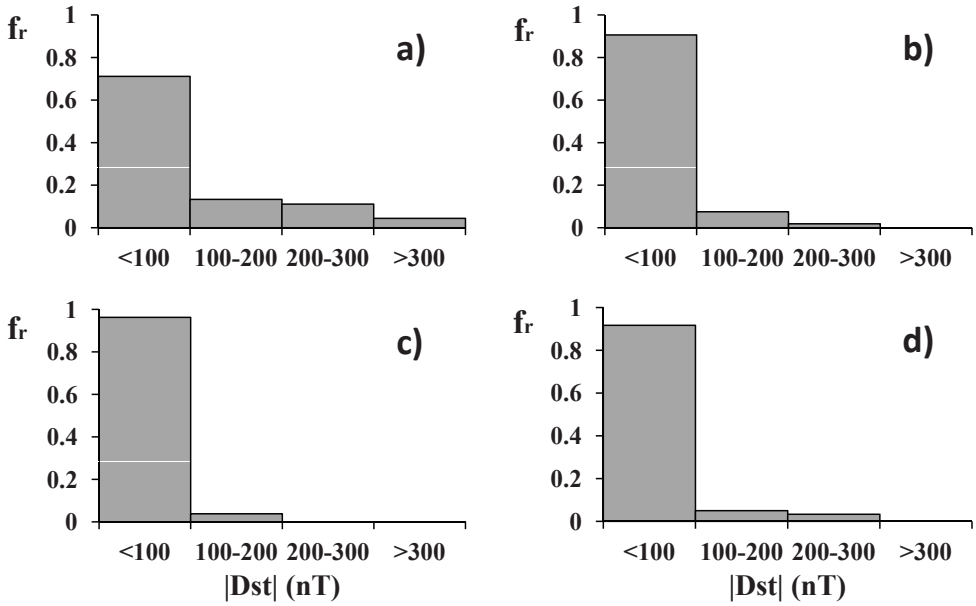

Figure 6. $|Dst|$ relative frequencies for different bins of the distance of the CME/flare source from the solar disc centre, r, expressed in the units of solar radius: a) $r < 0.4$; b) $0.4 < r < 0.6$; c) $0.6 < r < 0.8$; d) $r > 0.8$

Due to the fact that the CME–CME trains are considered as one entity (T? and T CMEs, as explained in Section 2), associated with a source position of the fastest event within the train, the complete analysis was repeated for S and S? samples (a total of 132 CMEs). All of the mentioned categories (quadrant, north/south, east/west and distance from the disc centre) show quite a similar behaviour.

Additionally, we inspected a dependence on the central meridian distance (CMD), *i.e.*, the distance of the CME/flare source position relative to the central meridian on the visible solar disc. The events were categorized as follows: $-90° <$ CMD $< -60°$, $-60° <$ CMD $< -30°$, $-30° <$ CMD $< 0°$, $0° <$ CMD $< 30°$, $30° <$ CMD $< 60°$, and $60° <$ CMD $< 90°$, with number of events per bin 26, 36, 40, 51, 36, and 22, respectively. A small E–W asymmetry can be observed for $30° <$ CMD $< 60°$, due to the fact that out of 36 east events in this bin, none had a $|Dst| > 100$ nT. This bin is also significantly different from all other bins. However, there is a lack of significant E–W differences between the rest of the CMD samples. Therefore, contrary to studies that report E–W asymmetry (*e.g.*, Zhang *et al.*, 2003; Zhang *et al.*, 2007; Gopalswamy, Yashiro, and Akiyama, 2007), our analysis shows more or less symmetrical longitudinal distribution of geoeffective CMEs in agreement with, *e.g.*, Srivastava and Venkatakrishnan (2004).

An alternative binning of the source position distance from the solar disc centre, r, was made, with the same purpose as in Section 4.1. The alternative r bins cover ranges: $r < 0.35$, $0.35 < r < 0.5$, $0.5 < r < 0.65$, $0.65 < r < 0.78$, $0.78 < r < 0.92$, and $r > 0.92$. The number of events in each bin is 35, 38, 37, 33,

Table 2. The significance results for the two sample t-test with equal variance not assumed, for the $|Dst|$ distribution mean, between different bins of: a) the source-location distance from the solar disc centre, r; b) interaction parameter. Unless marked with an asterisk, the value states that the means of the two samples are not significantly different; ** denotes that the significance of the difference is $> 95\%$; * denotes that the significance of the difference is $> 90\%$.

a) r bins			
	bin1[1]	bin2	bin3
bin4	0.002**	0.24	0.86
bin3	0.001**	0.10*	—
bin2	0.01**	—	—
b) interaction parameter			
	S[2]	S?	T?
T	0.001**	0.43	0.45
T?	0.06*	0.91	—
S?	0.12	—	—

[1]bins 1–4 represent different r ranges in units of solar radii: < 0.4 (bin1), 0.4-0.6 (bin2), 0.6-0.8 (bin3), and > 0.8 (bin4)
[2]different interaction parameters: no interaction (S), interaction not likely (S?), interaction probable (T?), and interaction highly probable (T)

37 and 31, respectively. The distribution mean for these alternative distributions are shown as gray dots in Figure 9a, together with the original binning (black dots). We can see that binning does not change the result notably, as both distributions follow the same trend.

4.3. CME–CME interaction parameter

As explained in Section 2 the CME–CME interaction parameter was defined employing four categories: SINGLE (S), SINGLE? (S?), TRAIN? (T?), and TRAIN (T). The number of events in each bin is 98, 34, 28, and 51, respectively. For each interaction parameter, a $|Dst|$ distribution was made, as in Sections 4.1 and 4.2. This resulted in four $|Dst|$ distributions shown in Figure 7. The results of the two sample t-test are presented in Table 2b.

It can be seen in Figure 7 that the distribution for "S-events" has a long tail, but is very asymmetric and shifted towards lower values of $|Dst|$. As the interaction level shifts from "S?" to "T?" and "T" the distribution "fills up" the tail and therefore shifts towards larger values of $|Dst|$. The results of the two sample t-test are somewhat inconclusive, because there is a significant difference only between "S" and "T" samples. However, we see that the probabilities that the two samples are statistically the same, decreases with the interaction level, and is highest for the neighbouring bins, thus implying that the effect comes from mixing the bins ("S?" and "T?" are actually mixtures of "S" and "T" events, with "S?" presumably dominated with "S" events and "T?" with "T" events, respectively). Therefore, we conclude that indeed CME–CME interaction

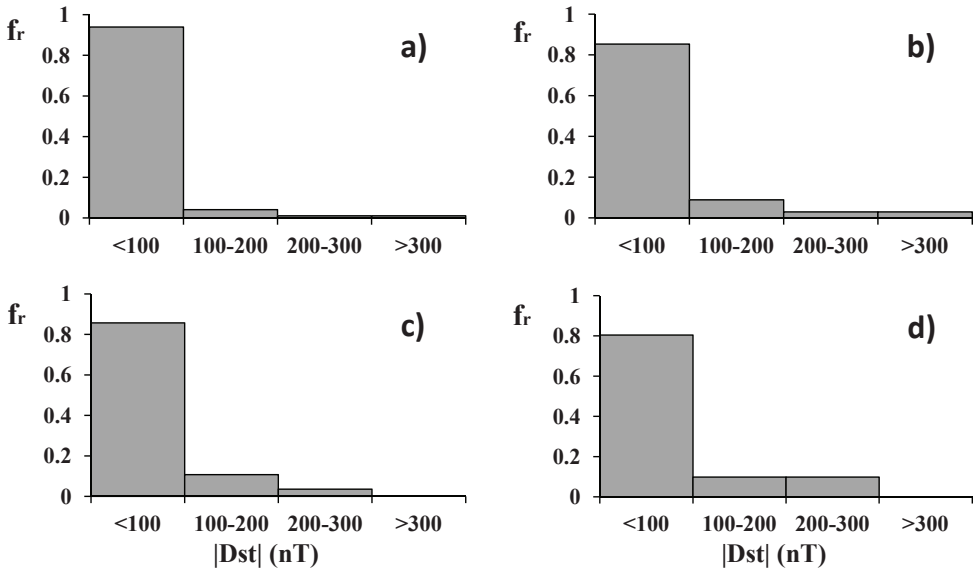

Figure 7. $|Dst|$ relative frequencies for different interaction levels: a) S events - no interaction; b) S? events - interaction not likely; c) T? events - interaction probable; d) T events - interaction highly probable.

influences the probability of a certain $|Dst|$ level, where we can associate higher probabilities of intense storms to CME trains. This does not mean that CME trains are more geoeffective due to some physical mechanism, because we also observe "S-CMEs" producing extremely intense ($|Dst| > 300$ nT) storms. Our results just show that they are less likely to do that.

However, due to the fact that CMEs in a train are regarded as one entity, with the CME parameters defined by that of the fastest CME in the train, the relationship between the interaction parameter and geoeffectiveness could simply be a byproduct of the relationship between CME speed and geoeffectiveness. Indeed, the speed distribution of "T" CMEs is shifted to larger speeds as opposed to "S" CMEs, *i.e.*, "T" CMEs are generally associated with larger 1$^{\rm st}$ order (linear) CME speed than "S" CMEs. On the other hand, when we exclude the fastest CMEs from the sample ($v >$1700 km s^{-1}) the difference in the speed distribution between "T" and "S" CMEs is lost, whereas the relationship between the interaction parameter and $|Dst|$ does not change notably. Therefore, although there is a relationship between the CME speed and the CME interaction parameter, it seems it is not the source of the relationship between the interaction parameter and geoeffectiveness.

To substantiate our results, we mixed the neighbouring bins (S with S?, S? with T?, and T? with T) and thus obtained three additional distributions. The distribution mean for the original interaction bins (black dots) and mixed interaction bins (gray dots) is plotted in Figure 9b, where numerical values were attributed to different interaction levels for quantification reasons ("S"=1; "S?"=2; "T?"=3; "T"=4). It can be seen that the mixed bins follow the same trend as

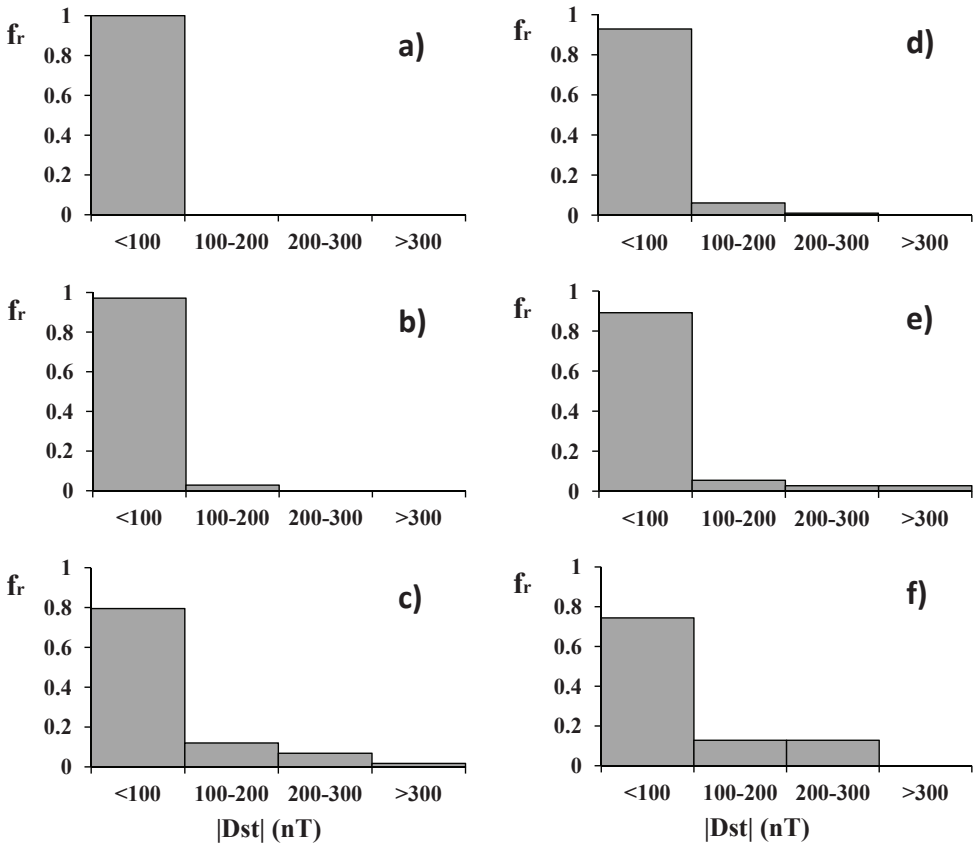

Figure 8. $|Dst|$ relative frequencies for different CME-width bins (a – c) and flare-class bins (d – f): a) non halo CMEs; b) partial halo CMEs; c) halo CMEs; d) B&C class flares; e) M class flares; f) X class flares;

original bins. A power-law function is fitted to the four levels of interaction to illustrate this trend.

4.4. CME angular width

The binning for CME (apparent) width, w, follows the categorization from the SOHO LASCO CME catalog into non-halo ($w < 120°$), partial halo ($120° < w < 360°$) and halo CMEs ($w = 360°$). Due to the fact that interacting CMEs are regarded as one entity (see Section 2), "T" and "T?" events were associated with the width of the widest CME within a train (*i.e.*, halo or partial halo, if present). The number of events within a certain width bin for non-halo, partial halo, and halo CMEs are 59, 35, and 117, respectively. Using $|Dst|$ binning explained previously, three $|Dst|$ distributions were made (Figures 8a–c). The results of the two sample t-test are presented in Table 3a.).

In Figure 8 we see an obvious progression in the $|Dst|$ distribution towards larger $|Dst|$ as the apparent width of the CME increases. For non-halos we find

Table 3. Two sample t-test significance levels for the $|Dst|$ distribution mean with equal variance not assumed for: a) different CME-width bins; b) different flare-class bins. Unless marked with an asterisk, the value states that the means of the two samples are not significantly different; ** denotes that the significance of the result is $> 95\%$; * denotes that the significance of the result is $> 90\%$

a) Width bins			b) Flare-class bins		
NH *vs* PH[1]	NH *vs* H	PH *vs* H	B&C *vs* M[2]	B&C *vs* X	M *vs* X
0.06*	$9 \cdot 10^{-10}$**	$9 \cdot 10^{-6}$**	0.17	10^{-3}**	0.03**

[1]NH = non-halo CMEs ($w < 120°$); PH = partial halo CMEs ($120° < w < 360°$); H = halo CMEs ($w = 360°$)
[2]B, C, M, X = B, C, M, X class flares

one-bin distribution within $|Dst| < 100$ nT, for partial halos the distribution gains a small tail, whereas for halos a long tail is observed. The distribution mean has an obvious increasing trend with larger widths (black dots in Figure 9c), which can be fitted by a quadratic function. These results are confirmed with the two sample t-test, showing that non-halo, partial halo and halo CME associated $|Dst|$ distributions are significantly different (Table 3a.).

The analysis was repeated separately for "S" and "S?" CMEs, with the same results and minor loss in significance (due to smaller number of events). Furthermore, we associate the width of the fastest CME in a train (as opposed to the previous association of the widest CME in a train) to "T" and "T?" events and repeat the analysis. Similar results are obtained. Both the distribution for non-halo and partial halo CMEs are restricted to $|Dst| < 200$ nT, but the distribution mean for partial halos is somewhat larger (although not statistically significant).

These results confirm a widely accepted view that halo CMEs are more geoeffective (*e.g.*, Zhang *et al.*, 2003; Srivastava and Venkatakrishnan, 2004; Gopalswamy, Yashiro, and Akiyama, 2007), as they clearly show that halo CMEs have larger probabilities to cause intense storms. In addition, we can conclude that non-halo CMEs are not likely to produce major storms ($|Dst| > 100$ nT) unless they are involved in a CME–CME interaction with a wider CME.

Finally, an alternative width binning was applied, as in previous sections (see Sections 4.1–4.3), using SOHO LASCO CME catalog table values for the apparent width. Again, "T" and "T?" events were associated with the width of the widest CME within a train. The alternative width bins are: $w < 70°$ (29 events), $70° < w < 130°$ (32 events), $130° < w < 360°$ (33 events), and $w = 360°$ (halo CMEs, 117 events). The distribution mean for the original width bins (black dots) and alternative width bins (gray dots) is plotted in Figure 9c. The numbers were associated to different width bins for quantitative reasons (non-halo CME=1; partial halo CME=2, halo CME=3). It can be seen that the alternative width bins follow the same trend as the original bins. A quadratic function is fitted to the original width bin data (non-halo, partial halo and halo CMEs) to illustrate this trend.

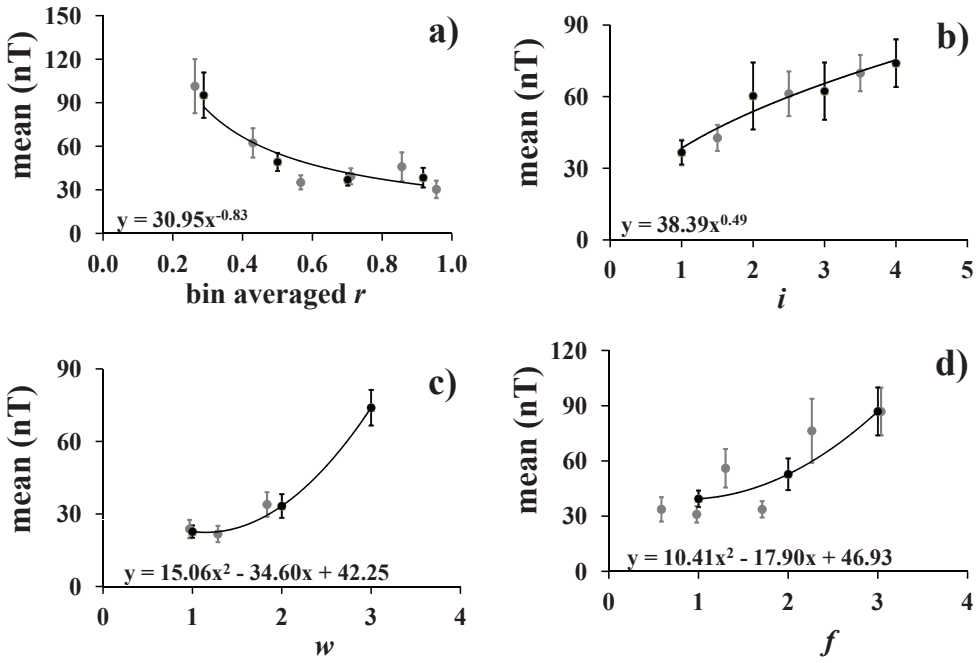

Figure 9. $|Dst|$ distribution mean as a function of: a) average value of the source position distance from the solar disc centre, r, within a specific bin; b) interaction parameter, i ("S"=1, "S?"=2, "T?"=3; "T"=4); c) width, w (non-halo=1, partial halo=2, halo=3); d) flare class, f (B&C=1, M=2, X=3). Black and gray dots mark different types of binning (for detailed explanation see Sections 4.2-4.5). Error bars represent confidence intervals, whereas the line shows the fitting curve (fitted trough black dots).

4.5. Flare X-ray class

The binning of the solar flare class follows the categorization of soft X-ray flares according to their soft X-ray flux peak value (F_{max} in units Wm^{-2}): $F_{max} < 10^{-6}$ (B class flare), $10^{-6} \leq F_{max} < 10^{-5}$ (C class flare), $10^{-5} \leq F_{max} < 10^{-4}$ (M class flare), and $F_{max} \geq 10^{-4}$ (X class flare). Due to lack of events associated with a B class flare, they were put in the same bin with C class flares. Therefore, binning into three flare categories was applied, namely B&C class flares, M flares, and X flares. The number of events are 98, 74, and 39, respectively. Three $|Dst|$ distributions were made (Figures 8d–f). We can observe minor differences in $|Dst|$ distribution between B&C class flares and M class flares, whereas there is a clear difference compared to X class flares distribution, which contains substantially larger fraction of events in its tail than other two distributions. This is also reflected in the two sample t-test, showing that although B&C class flares and M flares are not significantly different samples, they are both significantly different from X flares (Table 3).

The distribution mean has an increasing trend with flare class, which can be illustrated by a quadratic function, similar to that shown in Section 4.4 (Figure 9d). An alternative binning was again applied (gray dots) showing similar trend, however with a large scatter. For alternative binning, we used the peak value of

the X-ray flux (F_{max} in units Wm^{-2}) in the following ranges: $F_{max} < 2.5 \cdot 10^{-6}$ (31 event), $2.5 \cdot 10^{-6} \leq F_{max} < 5 \cdot 10^{-6}$ (40 events), $5 \cdot 10^{-6} \leq F_{max} < 1.2 \cdot 10^{-5}$ (34 events), $1.2 \cdot 10^{-5} \leq F_{max} < 3 \cdot 10^{-5}$ (35 events), $3 \cdot 10^{-5} \leq F_{max} < 10^{-4}$ (36 events), $F_{max} \geq 10^{-4}$ (35 events). The numbers were associated to different flare bins for quantitative reasons, similarly as in Sections 4.3 and 4.4 (B&C class=1, M class=2, X class=3).

The analysis was repeated for S and S? CMEs, with similar results, but with a loss in significance (only 19 X class flare events). Nevertheless, we can conclude that geoeffective CMEs are associated with stronger flares, in agreement with previous studies (e.g. Srivastava and Venkatakrishnan, 2004; Zhang *et al.*, 2007).

4.6. Combinations of solar parameters

To investigate the influence of combined solar parameters on the $|Dst|$ level, bivariate $|Dst|$ distributions for different combinations of two solar parameters were estimated. The same $|Dst|$ bins were used as in Sections 4.1–4.5. Table 4 shows the range and median values of $|Dst|$ distributions for different combinations of two solar parameters.

It can be seen in Table 4a that the median of the $|Dst|$ distribution has highest values for speed bins where $v > 1200$ km s^{-1} in case of non-halo and partial halo CMEs, whereas for halo CMEs this is the case when $v > 1700$ km s^{-1}. Therefore, the combination of a larger apparent width and larger speed of CMEs increases the probability of a larger $|Dst|$ level (*i.e.*, stronger geomagnetic storm). It should be noted that the median value for halo CMEs in almost all of the speed bins is larger compared to non-halo and partial-halo CMEs. This again indicates the importance of the apparent width as a relevant solar parameter regarding geoeffectiveness, in agreement with previous studies (*e.g.*, Zhang *et al.*, 2003; Srivastava and Venkatakrishnan, 2004; Gopalswamy, Yashiro, and Akiyama, 2007).

In Table 4b the combination of CME speed, v, and source distance from the solar disc centre, r, is investigated. It can be seen that the central source positions are associated with higher values of the $|Dst|$ distribution median, as well as larger speeds. Furthermore, combination of a very fast CME ($v > 1200$ km s^{-1}) and a source position close to the disc centre have significantly higher median, *i.e.*, highly increases the chance of a strong geomagnetic storm. For the source positions closer to the limb we also observe that the CME speed plays a role in causing higher $|Dst|$ levels. We also observe a change in the median value of the $|Dst|$ distribution for the combination of very fast CMEs and very high interaction level (Table 4c). However, this is not so pronounced as in the case of the speed/source-position combination. Finally, the width/source-position combination of solar parameters leads to the highest $|Dst|$ levels for halo CME from the disc centre (Table 4d) although halo CMEs closer to the limb can also cause higher $|Dst|$ values.

From the presented analysis we conclude that the combination of favorable solar parameters can result in enhanced geoeffectiveness. These model parameters individually do not have the same impact. The latter is also visible from the fitted curves in Figures 5 and 9: the fitted functions for different solar parameters have different growth/descend rates. A combination of solar parameters

Table 4. Minimum/maximum/median of the $|Dst|$ distribution for the parameter combinations: a) CME speed, v, and apparent width, w; b) CME speed, v, and source position distance from the solar disc centre, r; c) CME speed, v, and interaction parameter; d) apparent width, w, and source position distance from the solar disc centre, r. Values are not displayed whenever there are five or less events included.

a) *v versus w*

v (km s^{-1})	NH[1]	PH	H
400-600	0/80/20	—	10/60/40
600-800	0/90/20	0/80/30	10/190/45
800-1000	0/60/20	0/55/35	0/380/50
1000-1200	0/30/10	0/40/30	0/280/40
1200-1700	20/190/30	0/140/40	0/410/35
>1700	—	—	0/280/70

b) *v versus r*

v (km s^{-1})	$r < 0.4$	$0.4 < r < 0.6$	$0.6 < r < 0.8$	$r > 0.8$
400-600	0/110/50	—	0/75/20	10/30/10
600-800	0/190/33	0/80/30	15/90/40	0/50/20
800-1000	0/380/38	10/100/40	10/70/40	0/40/10
1000-1200	—	0/140/40	0/30/10	0/280/30
1200-1700	0/410/215	0/150/35	0/140/33	0/190/30
>1700	40/280/130	30/250/85	0/90/45	10/210/50

c) *v versus* interaction parameter

v (km s^{-1})	S[2]	S?	T?	T
400-600	0/90/20	—	—	—
600-800	0/190/20	—	10/90/30	15/90/45
800-1000	0/380/30	10/100/55	—	0/150/33
1000-1200	0/140/30	0/280/15	—	20/130/30
1200-1700	0/140/30	0/280/20	20/270/150	10/250/40
>1700	0/270/30	20/140/85	—	60/280/90

d) *w versus r*

width	$r < 0.4$	$0.4 < r < 0.6$	$0.6 < r < 0.8$	$r > 0.8$
NH	0/30/10	0/80/30	0/90/20	0/190/20
PH	30/110/45	25/30/28	0/140/45	0/75/15
H	0/410/65	0/250/40	0/110/35	0/280/55

[1]NH = non-halo CMEs ($w < 120°$); PH = partial halo CMEs ($120° < w < 360°$); H = halo CMEs ($w = 360°$)
[2]for detailed explanation see Section 2

for investigating geoeffectiveness has been attempted by Srivastava (2005) and Srivastava (2006). However, her conclusion was that without the interplanetary parameters, the forecast of CME geoeffectiveness is insufficiently precise.

4.7. Monthly CME/flare and *Dst* activity

Finally, we investigate the monthly CME/flare activity in the SOHO era. For that purpose we define monthly CME/flare activity parameter in a given month, $A_{\rm i}$ (i=1,2,...,12), based on the solar parameters that influence the geoeffectiveness, as derived throughout Sections 4.1–4.5:

- number of CMEs in month "i" of the year, $N_{\rm CME,i}$
- average CME speed in month "i", $v_{\rm avg,i}$
- number of CMEs with speed > 1000 km s^{-1} in month "i", $N_{v,{\rm i}}$
- number of X-class flares in month "i", $N_{\rm X,i}$
- number of HALO CMEs in month "i", $N_{\rm HALO,i}$
- average source position distance from the solar disc centre of CMEs/flares in month "i", $r_{\rm avg,i}$ (in units of solar radii)
- SOHO downtime in hours in month "i", $t_{\rm SOHO,i}$

These parameters were chosen in a way that they not only represent the "quality" of events (*i.e.*, connection to geoeffectiveness) but also the occurrence rate of events. We also include the SOHO downtime as the parameter to account for possible lack of important CME data. Each of these parameters was ranked, *i.e.*, associated with an ordinal 1 – 12, depending on the value it obtained for each month (see Table 5). The ranks were associated so that the highest rank (*i.e.* lowest ordinal) roughly corresponds to highest geoeffectiveness.

The monthly CME–flare activity parameter in month "i" (i=1,2,...,12), $A_{\rm i}$, is defined as follows:

$$A_{\rm i} = N_{{\rm CME},i} + v_{{\rm avg},i} + N_{v,i} + N_{{\rm X},i} + N_{{\rm HALO},i} + r_{{\rm avg},i} + t_{{\rm SOHO},i}. \qquad (1)$$

The monthly CME/flare activity parameter was normalized to obtain the relative monthly CME/flare activity parameter (in the month i=1,2,...12), $A_{\rm rel,i}$:

$$A_{\rm rel,i} = \frac{A_{\rm i}}{\sum_{i=1}^{12} A_{\rm i}}. \qquad (2)$$

The relative monthly CME/flare activity parameter was obtained both for the 1392 CME–flare pairs in the SOHO era and the 211 *Dst*-associated CME–flare pairs used for the statistical analysis throughout Sections 4.1–4.6. The two are compared in Figure 10a, where it can be seen that they have a very similar trend. Furthermore, we examined their variations, δA, *i.e.*, the residuals when the two are subtracted (Figure 10b). Variations of the two curves, δA, are distributed in a normal-like distribution centred around ≈ 0, therefore, we conclude that the two curves indeed have the same trend. This means that the CME/flare activity of our sample (211 *Dst*-associated CMEs) is a good representative of the population (1392 CME–flare pairs in the SOHO era).

Table 5. Values and ranks of solar parameters in a specific month for 1392 CME–flare pairs in the SOHO era. Definitions of solar parameters are given in the main text.

	Solar parameter value						
Month, i	$N_{\mathrm{CME,i}}$	$v_{\mathrm{avg,i}}$(km s^{-1})	$N_{v,\mathrm{i}}$	$N_{\mathrm{X,i}}$	$N_{\mathrm{HALO,i}}$	$r_{\mathrm{avg,i}}$	$t_{\mathrm{SOHO,i}}$(h)
1	90	621	13	4	18	0.6828	1456
2	69	543	6	2	11	0.7170	1740
3	99	552	9	3	11	0.7334	1232
4	137	659	21	9	20	0.7058	999
5	133	593	17	5	24	0.7028	808
6	130	604	15	4	18	0.7610	3665
7	142	585	19	14	23	0.7583	825
8	133	559	15	10	13	0.7276	839
9	90	672	17	4	14	0.6460	1081
10	131	586	18	10	16	0.6648	500
11	139	682	27	16	36	0.6441	1680
12	99	573	11	5	21	0.7249	1745
	Solar parameter rank						
Month, i	$N_{\mathrm{CME,i}}$	$v_{\mathrm{avg,i}}$	$N_{v,\mathrm{i}}$	$N_{\mathrm{X,i}}$	$N_{\mathrm{HALO,i}}$	$r_{\mathrm{avg,i}}$	$t_{\mathrm{SOHO,i}}$
1	2	9	4	3	6	9	8
2	1	1	1	1	1	6	10
3	4	2	2	2	2	3	7
4	10	10	11	8	8	7	5
5	8	7	7	6	11	8	2
6	6	8	5	4	7	1	12
7	12	5	10	11	10	2	3
8	9	3	6	9	3	4	4
9	3	11	8	5	4	11	6
10	7	6	9	10	5	10	1
11	11	12	12	12	12	12	9
12	5	4	3	7	9	5	11

Then, using 211 *Dst*-associated CME–flare pairs, monthly *Dst* activity parameter was obtained using the following parameters:

- number of events in month "i" of the year, N_{i}
- average $|Dst|$ values for month "i", $|Dst|_{\mathrm{avg,i}}$

Again, both qualitative and quantitative aspects were taken into account. The number of events, N_{i}, was ranked by the value in each month from 1 to 12, where 1 was associated to the month where there was the smallest number of events, and 12 was associated to the month where there was the largest number of events. Similarly, the average values, $|Dst|_{\mathrm{avg,i}}$, were also ranked, where 1 was associated to the month where $|Dst|_{\mathrm{avg,i}}$ assumes the lowest value and 12 was associated to the month where $|Dst|_{\mathrm{avg,i}}$ assumes the highest value. Using *Dst* parameter ranks, monthly *Dst* activity parameter, A_{i}, for specific month i (i=1,2,...,12) was obtained:

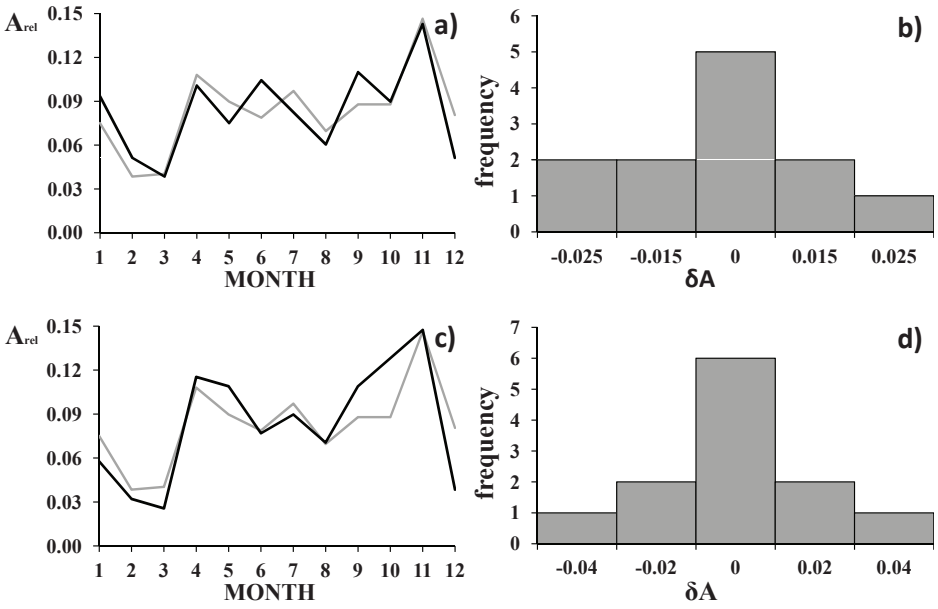

Figure 10. a) Relative monthly CME/flare activity parameter ($A_{\rm rel}$) derived from the 1392 CME–flare pairs in the SOHO era (gray) and from the 211 *Dst*-associated CME–flare pairs (black); b) distribution of variations, δA, between curves displayed in Figure 10a. The mean value is $3.5 \cdot 10^{-18}$ and standard deviation is 0.017; c) relative monthly CME/flare activity parameter ($A_{\rm rel}$) derived from the 1392 CME–flare pairs in the SOHO era (gray) and relative monthly *Dst* activity parameter obtained from the 211 *Dst*-associated CME–flare pairs (black); d) distribution of variations, δA, between curves shown in Figure 10c. Mean value is $2.3 \cdot 10^{-18}$ and standard deviation is 0.021.

$$A_{\rm i} = N_{\rm i} + |Dst|_{\rm avg,i}. \tag{3}$$

The monthly *Dst* activity parameter was normalized to obtain the relative monthly *Dst* activity parameter (in the month i=1,2,...12) using Equation (2). We compare it to the relative monthly CME/flare activity parameter in Figure 10c, and again we find that the two have a very similar trend. The variations, δA, of the two curves (Figure 10d) are distributed in a normal-like distribution centred around ≈ 0, similarly as in Figure 10b. Therefore, we reach the same conclusion, that the two curves have the same trend. This means that the monthly *Dst* activity, derived from our sample reflects the monthly CME/flare activity in the SOHO era.

5. Empirical statistical model for predicting geomagnetic storm levels

In Sections 4.1–4.5 the key solar parameters were examined and were found to be related to the $|Dst|$ levels. Namely, the distribution of $|Dst|$ amplitudes was found to change with CME speed, v, CME/flare source region location (distance

from the centre of the solar disc, r), CME apparent width, w, flare class, f, and CME–CME interaction level, i. These relationships are quantified and can be used to predict the probability for the $|Dst|$ levels based on the remote solar observations. We use the results of the statistical analysis to construct $|Dst|$ distributions, depending on the specific solar parameter. For this purpose we use the geometric distribution:

$$P(X = k) = p \cdot (1 - p)^{k-1}\,, \tag{4}$$

where $P(X = k)$ is the probability that the k^{th} trial is a first success and p is the probability of the success in each trial ($k = 1, 2, 3, ...$ is the number of trials). The geometric distribution is suitable for several reasons. It is a rapidly descending discrete distribution, like the $|Dst|$ distribution we observe, and therefore is restricted to a small number of bins. In addition, it can be simply mathematically reconstructed based on the distribution mean ($p = m^{-1}$, where m is the distribution mean). The association between the number of trials and the $|Dst|$ bins was made in the following way:

- $k = 1 \longleftrightarrow |Dst| < 100\,\text{nT}$;
- $k = 2 \longleftrightarrow 100\,\text{nT} < |Dst| < 200\,\text{nT}$;
- $k = 3 \longleftrightarrow 200\,\text{nT} < |Dst| < 300\,\text{nT}$;
- $k = 4 \longleftrightarrow |Dst| > 300\,\text{nT}$.

In this way, the conversion of the $|Dst|$ distribution mean, m_{DST}, into the geometric distribution mean, m_{GD}, can be done in a simple way ($m_{GD} = 1 + m_{DST}\,[\text{nT}]/100$, for details see Appendix).

It was shown in Figures 5 and 9 that the trend of the change in the $|Dst|$ distribution mean, m_{DST}, with a specific solar parameter can be fitted by a corresponding function. Namely, $m_{DST}(v)$ was fitted with a linear function, $m_{DST}(r)$ and $m_{DST}(i)$ with a power-law function, and $m_{DST}(w)$ and $m_{DST}(f)$ with a quadratic function. Therefore, based on the $|Dst|$-solar parameter relationships found, a corresponding geometric distribution can be obtained. We note that the CME speed, v, and the CME source distance from the centre of the solar disc, r, are regarded as continuous parameters in the ranges of $v \geq 400\ \text{km}\,\text{s}^{-1}$ and $0 < r \leq 1$, respectively. The range of v is determined based on the limitations of the sample, whereas the range of r is restricted by the mathematical singularity of the power-law function ($r = 0$) and the physical boundary ($r = 1$, *i.e.*, the solar limb). The other three solar parameters, the apparent width, w, the associated flare class, f, and the level of interaction, i, are considered as discrete parameters associated with integers 1–3 and 1–4, respectively (1 meaning least significant, *i.e.*, the lowest interaction parameter, width, and flare class).

The mathematically obtained geometric distribution underestimates the observed $|Dst|$ distribution for $k = 1$, whereas it is overestimated for $k = 2$. This can be seen in Figure 11, where the two are compared for a number of relationships. Therefore, new "adjusted" distributions for each of the key solar parameters were obtained by adding a specific constant to each bin to best fit the observed distribution in all the ranges, *i.e.*, for all the values of key

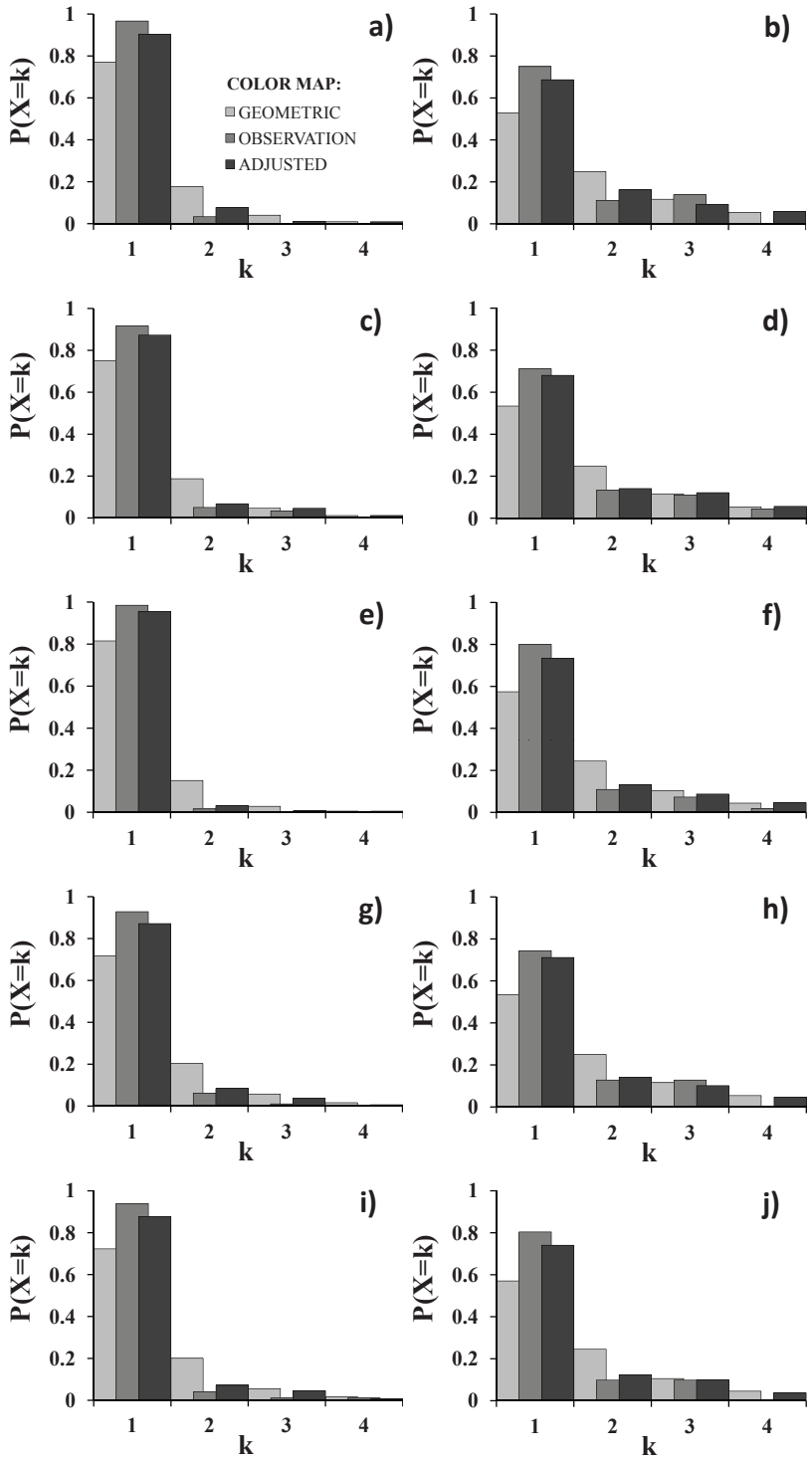

Figure 11. Geometric, observational and adjusted distributions for different ranges/values of key solar parameters: a) for the CME speed, 400 $\mathrm{km\,s^{-1}}$< v <600 $\mathrm{km\,s^{-1}}$; b) for the CME speed, v >1700 $\mathrm{km\,s^{-1}}$; c) for the CME source distance from the centre of the solar disc, $r > 0.8$; d) for the CME source distance from the centre of the solar disc, $r < 0.4$; e) for non-halo CMEs, $w < 120°$ ($w = 1$); f) for halo CMEs, $w = 360°$ ($w = 3$); g) for the associated flares of B&C-class, $f = 1$; h) for the associated flares of X-class, $f = 3$; i) for the lowest interaction parameter, "S" (SINGLE), $i = 1$; j) for the highest interaction parameter, "T" (TRAIN), $i = 4$.

Table 6. The constants added to geometric distribution to obtain adjusted distribution, for different $|Dst|$ bins, k, and different solar parameters.

k	v[1]	r	w	f	i
1	0.13	0.12	0.14	0.15	0.15
2	-0.10	-0.12	-0.12	-0.12	-0.13
3	-0.03	0	-0.02	-0.02	-0.01
4	0	0	0	-0.01	-0.01

[1]solar parameters: CME speed, v, CME source position distance from the centre of the solar disc, r, apparent width, w, associated flare class, f, and interaction parameter, i

solar parameters. The constants are added so that the new distribution is also normalized (Table 6) and are different for different $|Dst|$ bins, *i.e.*, k and different key solar parameters. It can be seen in Figure 11 that the empirical distribution still slightly underestimates the observed $|Dst|$ distribution for $k = 1$. However, the agreement between the two distributions for higher values of k is substantially improved. For detailed mathematical formulations and procedures used to obtain probability distributions see Appendix.

The obtained empirical distributions are treated as probability distributions. For a specific solar parameter they provide the information on the probability for associating it a specific value of k, *i.e.*, $|Dst|$ level. To combine the effect of the key solar parameters, *i.e.*, to obtain a joint probability distribution, the key parameters were treated as mutually non-exclusive events, for which the following formula applies:

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A} \cap \mathrm{B}). \tag{5}$$

In general, $P(\mathrm{X})$, where $X = A, B$, is the (marginal) probability of the event X, $P(\mathrm{A} \cup \mathrm{B})$ is the probability that either event A or event B or both occur, and $P(\mathrm{A} \cap \mathrm{B})$ is their joint probability. Specifically, in our case $P(X) = P(X = k)$ is the probability that for a specific key solar parameter X a specific $|Dst|$ level, k, will be observed. It should be noted that since a particular key parameters are tied to the same event, they should be regarded as mutually non-exclusive. In general, joint probability is given by:

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}|\mathrm{B}) \cdot P(\mathrm{B}). \tag{6}$$

Where $P(\mathrm{A}|\mathrm{B})$ is the conditional probability, *i.e.* the probability for event A given that the event B occured. Assuming that the events are independent of each other, combining Equations (5) and (6) one gets:

$$P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A}) \cdot P(\mathrm{B}). \tag{7}$$

This assumption is not fully valid, due to the fact that not all key solar parameters are independent of each other, (*e.g.*, CME speed and flare class, see Moon *et al.*, 2002, Moon *et al.*, 2003, Vršnak, Sudar, and Ruždjak, 2005, Maričić *et al.*, 2007). Since the constructed geometric distribution directly depends on the key solar parameter for which it is constructed, the connection between

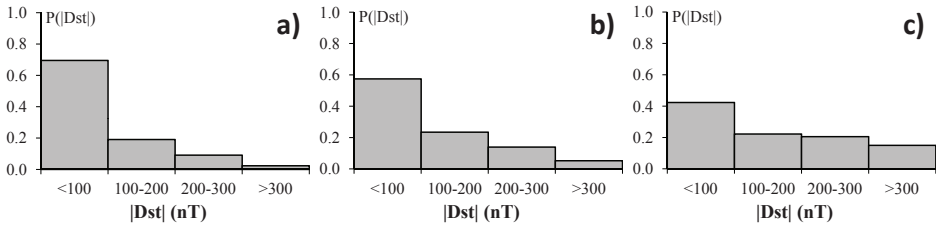

Figure 12. Probability distribution for observing $|Dst|$ in a specific $|Dst|$ bin for different sets of key solar parameters: a) $v = 400$ km/s; $r = 1$; $w = 1$; $f = 1$; $i = 1$; b) $v = 800$ km/s; $r = 0.5$; $w = 2$; $f = 2$; $i = 2$; c) $v = 2000$ km/s; $r = 0.01$; $w = 3$; $f = 3$; $i = 4$.

two solar parameters leads to a relationship between two constructed geometric distributions. Moreover, positively correlated parameters will lead to conditional probability greater than the marginal probability for k=2,3,4 and *vice versa* for k=1. Consequently, the assumption of independence redefines parameter space in a way that it will at worst underestimate the joint probability $P(\mathrm{A} \cap \mathrm{B})$, *i.e.*, overestimate the probability $P(\mathrm{A} \cup \mathrm{B})$ for k=2,3,4 and *vice versa* for k=1. Therefore, the constructed probability distribution will (slightly) overestimate geoeffectiveness, increasing to some extent the number of false alarms.

Finally, the probability of observing the $|Dst|$ value in a specific bin for a set of solar key parameters is then given by the formula derived from Equation (7):

$$P(|Dst| = k) = \sum_{\alpha} P_\alpha - \sum_{\alpha \neq \beta} P_\alpha \cdot P_\beta + \sum_{\alpha \neq \beta \neq \gamma} P_\alpha \cdot P_\beta \cdot P_\gamma - \\ - \sum_{\alpha \neq \beta \neq \gamma \neq \delta} P_\alpha \cdot P_\beta \cdot P_\gamma \cdot P_\delta + \sum_{\alpha \neq \beta \neq \gamma \neq \delta \neq \epsilon} P_\alpha \cdot P_\beta \cdot P_\gamma \cdot P_\delta \cdot P_\epsilon \,, \tag{8}$$

where P_α=$P(\alpha)$ represents the probability of a $|Dst|$ level k for a specific solar key parameter α (CME speed, v, CME/flare source position distance from the centre of the solar disc, r, CME apparent width, w, flare class, f, and interaction parameter, i).

Based on the Equation (8), probabilities of $|Dst|$ levels can be calculated for a specific set of key parameters v, r, w, f, and i. In Figure 12 we present three different probability distributions obtained using Equation (8) for three different key solar parameter sets. The results are also presented in Table 7. It can be seen that the probabilities of large geomagnetic storms are higher for faster and wider CMEs which originate near the disc centre, are connected to more energetic flares and are likely to be involved in a CME–CME interaction.

This model constructs the geoeffectiveness probability distribution for a given CME. However, in its current form it cannot be used for forecasting. Since it is based on the distribution of CMEs in the SOHO era (a representative sample of 211 events) this probability distribution represents an ensemble of possible $|Dst|$ values for a given CME. Although the probability distribution changes with CME/flare parameters it is always highly asymmetric with greatest probability that CME will not be geoeffective. This depicts the general behavior of CMEs

Table 7. Probabilities for observing a specific $|Dst|$ level for different sets of key solar parameters: I: $v = 400$ km/s; $r = 1\mathrm{R_{SUN}}$; $w = 1$; $f = 1$; $i = 1$; II: $v = 800$ km/s; $r = 0.5\mathrm{R_{SUN}}$; $w = 2$; $f = 2$; $i = 2$; III: $v = 2000$ km/s; $r = 0.01\mathrm{R_{SUN}}$; $w = 3$; $f = 3$; $i = 4$.

$\lvert Dst\rvert$ (nT)	$P(\lvert Dst\rvert)(\%)$		
	I	II	III
<100	70	57	42
100-200	19	23	22
200-300	9	14	21
>100	30	43	58
>200	11	19	35
>300	2	5	15

- a large majority of CMEs will never reach the Earth and/or will not have a favorable magnetic field orientation. Therefore, although the model produces a probability distribution it does not give a straightforward prediction of whether or not (and how strong) a geomagnetic storm will occur.

This is a different approach than used in previous models (*e.g.* Srivastava, 2005; Valach *et al.*, 2009; Uwamahoro, McKinnell, and Habarulema, 2012), where prediction of geoeffectiveness is a direct output of the model. In Srivastava (2005) and Uwamahoro, McKinnell, and Habarulema (2012) the threshold for the probability function is set to 0.5, *i.e.* the prediction of the storm is based on the highest calculated probability. Here, one cannot simply predict the $|Dst|$ level by stating that it has the largest probability, because the largest probability for all the CMEs is that they will not produce $|Dst| > 100$ nT. In order to derive the forecast based on this model, one will have to impose thresholds on the probability distribution, similar to that done by Valach *et al.* (2009). This, however, requires further study and will be reported elsewhere.

The differences between this model and other models mentioned above arise from the basic sample choice: our sample is based on CMEs, whereas in other studies samples were based on ICMEs or geomagnetic storms (*i.e.* they presume the arrival of the CME at the Earth). In that sense, as explained in Section 3 our model also indirectly takes into account false alarms, because they are incorporated into the distribution. That does not mean that it can avoid false alarms completely. Due to assumption of independence leading to Equation (7), the effect of the false alarms may be increased, but this cannot be assessed at this point. In the present form the model is not suitable for space weather forecast and additional calculations are needed to derive the expected $|Dst|$ level for a specific probability distribution. Once this is achieved and evaluation performed, results can be properly compared to other models and false alarms studied in more detail.

6. Summary and Conclusions

From the presented statistical analysis we derived the key CME/flare parameters and quantified their influence on the probability of occurence of moderate and intense storms. Our results reflect some of relatively well-known relationships between remotely-observed solar properties and geomagnetic storms, namely the importance of CME initial speed, apparent width, source position and associated solar flare class. It also offers a quantification of these relationships and points out the significance in combining different solar parameters. It is shown that the CME–CME interaction is associated with a higher probability in causing intense storms. Moreover, it was shown that very slow, non-halo CMEs associated with B or C class flare are not expected to produce intense storms, unless involved in the CME–CME interaction with faster and wider CMEs, associated with stronger flares. The validity of the sample and of the results is confirmed by comparing the monthly CME/flare activity of the population (1392 CME–flare pairs in the SOHO era) with the monthly CME/flare and Dst activity in our sample (211 Dst-associated CME–flare pairs).

The results of this statistical analysis can be used for prediction of the probability that a given event, observed by coronagraph, X-ray and EUV imagers at L1-point satellites (or even ground based instruments), will produce a major, intense or very intense geomagnetic storm at the Earth. An empirical statistical model for predicting geomagnetic storm levels was established that can be used as an early geomagnetic storm warning. It calculates the $|Dst|$ level probability distribution for a set of key solar parameters, based on the discrete probability distribution constructed by means of a geometric distribution. The distribution is shifted towards larger $|Dst|$ levels for faster and wider CMEs which originate near the centre of the disc, especially if they are connected to more energetic flares and are likely to be involved in a CME–CME interaction. However, the distribution is always highly asymmetric with the highest probability that a CME will not be geoeffective, reflecting the general behavior of CMEs (majority of them never reach the Earth and/or do not have favorable magnetic field orientation). Therefore, in order to forecast based on this model, further analysis is needed. The prediction at this stage is quite "crude" and does not provide a straightforward information whether or not a geomagnetic storm will occur and what would be its intensity. However, its advantage is that it offers an advance warning.

It should be noted that non-geoeffective CMEs were basically treated equally as the CMEs which never reached the Earth. This is due to the sampling, where the general idea was to observe solar sources and effects at the Earth, without interplanetary component. We found this view to be appropriate regarding the statistical aspect, where physical variables are treated as random variables. Although the false alarms are included in the sample which has been used for developing the model, it is not possible to distinguish whether or not they occur. This might represent a drawback of the model which will be analyzed in a future work.

Appendix

Hereafter follows a supplement to Section 5, providing detailed step-wise mathematical formulations and procedures used for estimating the probability distribution of geomagnetic storm level based on the remote solar observation of a CME and the associated solar flare. For this purpose an example-CME will be used with the following characteristics:

- First LASCO C2 appearance: 10 April 2001 05:30 UT.
- Associated solar flare GOES peak time: 10 April 2001 05:26 UT.
- First order LASCO catalog CME speed: 2411 km s^{-1}.
- CME angular width: halo.
- CME/flare source region location (distance from the centre of the solar disc, r): S23W09 (r=0.4067).
- Flare X-ray class: X2.3.
- CME–CME interaction level: train, T (very likely interacts with a halo CME that first appeared in the C2 field of view 9 April 2001 15:54 UT).

Based on the CME/flare characteristics the key parameters are defined in the following way:

- v is a continuous parameter that is equal to the first order CME speed, measured in the LASCO field of view, expressed in km s^{-1} and defined in a range $v > 400$. Therefore, $v = 2411$.
- r is a continuous parameter that is equal to the distance of a CME/flare position from the centre of the solar disc expressed in solar radii and defined in a range $0 < r \leq 1$. Therefore, $r = 0.4067$.
- w is a discrete parameter with possible values 1 (non-halo CMEs), 2 (partial halo CMEs), and 3 (halo CMEs). Therefore, $w = 3$.
- f is a discrete parameter with possible values 1 (B or C flare), 2 (M flare), and 3 (X flare). Therefore, $f = 3$.
- i is a discrete parameter with possible values 1 (S, no interaction), 2 (S?, interaction not likely), 3 (T?, probable interaction), and 4 (T, interaction highly probable). Therefore, $i = 4$.

A. The geometric probability distribution, $P(X = k)$

The geometric probability distribution is a probability distribution of a random variable X, where X is a number of Bernoulli trials needed to get a success. There is an equal probability of success of each trial, p, and X is defined on an endless set of discrete values $k = 1, 2, 3, ...$ (see e.g. Pitman, 1993, and Stirzaker, 2003). It is a discrete analogue of the exponential distribution. The probability density function for the geometric distribution is given by Equation (4) in Section 5 and an example is given in Figure 13 for different probabilities of success in each trial, p.

It is easily found that the expected value of the geometrically distributed random variable X, i.e. the mean of the geometric distribution, is given by the following expression (for details see Stirzaker, 2003):

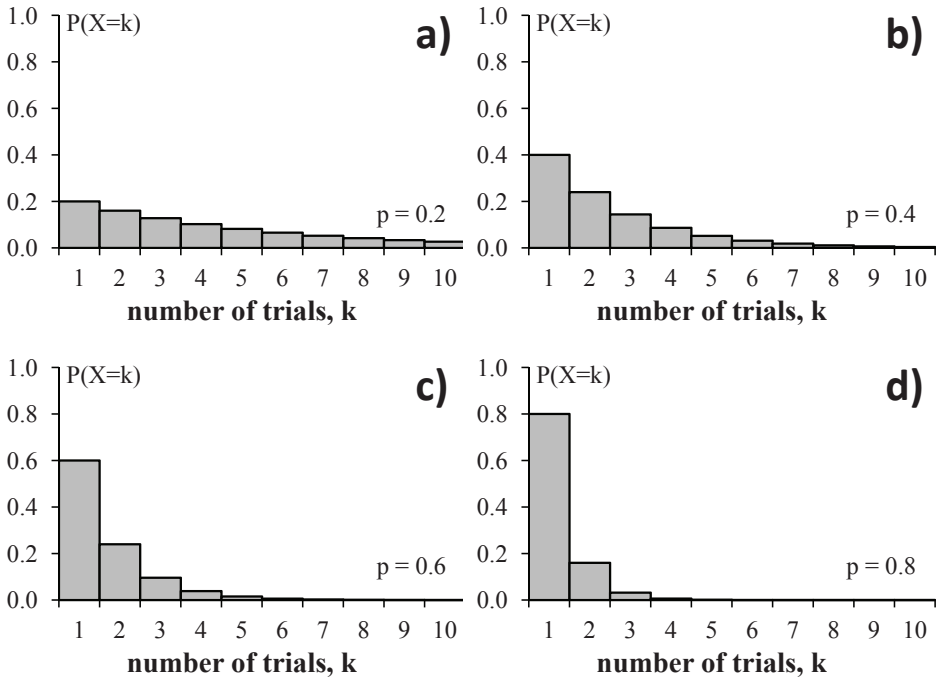

Figure 13. The geometric probability distribution, $P(X = k)$ for different probabilities of success on each trial, p.

$$m_{GD} = E(X) = \sum_{X \epsilon k} X \cdot P(X) = \sum_{k=1}^{\infty} k \cdot p \cdot (1-p)^{k-1} = \frac{1}{p}. \tag{9}$$

Therefore, the probability of the success in each trial, p can be calculated if the mean of the geometric distribution, m_{GD} is known:

$$p = \frac{1}{m_{GD}}. \tag{10}$$

We use the formalism for geometric distribution to construct $|Dst|$ distributions observed throughout Section 4. For that purpose, different $|Dst|$ levels have to be associated with different numbers of trials ($k \longleftrightarrow |Dst|$) and the mean of $|Dst|$ distribution has to be associated with the mean of the geometric distribution ($m_{GD} \longleftrightarrow m_{DST}$). We associate different $|Dst|$ levels with different number of trials, k in the geometric distribution in the following way:

- $k = 1 \longleftrightarrow |Dst| < 100\,\mathrm{nT}$;
- $k = 2 \longleftrightarrow 100\,\mathrm{nT} < |Dst| < 200\,\mathrm{nT}$;
- $k = 3 \longleftrightarrow 200\,\mathrm{nT} < |Dst| < 300\,\mathrm{nT}$;
- $k = 4 \longleftrightarrow |Dst| > 300\,\mathrm{nT}$.

Note that the value of k is exactly 100 times smaller than the upper boundary for associated $|Dst|$ level, expressed in nT. It would be reasonable to assume

that the mean of the geometric distributions would relate in a similar fashion to the mean of the $|Dst|$ distribution (*i.e.* would be 100 times smaller). The mean of the $|Dst|$ distribution is in the first bin for all of the distributions troughout Section 4, *i.e.* $m_{DST} < 100$nT. However, due to the fact that the geometric distribution is defined on a set $k = 1, 2, 3, ...$, the mean is always larger than 1. This is also seen from Equation 9 ($p < 1$). Dividing m_{DST} by 100 would not give a mathematically correct m_{GD}, but adding 1 to this relation solves this problem. Therefore:

$$m_{GD} = 1 + \frac{m_{DST}[nT]}{100} \tag{11}$$

For simplicity, in further reading we will refer to $P(X = k)$ as $P(k)$. Note that for constructed distributions we will use the set of k values $k = 1, 2, 3, 4$, based on the defined associations $k \longleftrightarrow |Dst|$.

B. Probability distribution for CME speed (v), $P_v(k)$

The change of the $|Dst|$ distribution mean with the CME speed, v, can be described with a linear function (see Figure 5):

$$m_{DST}(v) = a \cdot v + b. \tag{12}$$

Here $a = 0.04$, $b = 10.45$, and $|Dst|$ is expressed in nT. For a given $v = 2411$ Equation (12) gives distribution mean $m_{DST}(v) = 106.89$. The geometric distribution mean is then calculated using Equation (11), which in our example gives $m_{GD} = 2.07$. The probability of the success in each trial, p can be calculated using Equation (10) and in our example equals $p = 0.48$.

For each $k = 1, 2, 3, 4$, a probability that the k-th trial is the first success, $P(k)$ can be calculated using Equation (4) in Section 5. In the example the results are as follows:

- $P(k = 1) = 0.4833$;
- $P(k = 2) = 0.2497$;
- $P(k = 3) = 0.1290$;
- $P(k = 4) = 0.0667$.

Due to the fact that the geometric distribution does not stop at $k = 4$, this distribution is not normalized, *i.e.* $\sum_{k=1}^{4} P(k) \neq 1$. Therefore, to define this distribution for $k = 1, 2, 3, 4$, it is necessary to renormalize the distribution:

$$P(k) = \frac{P(k)}{\sum_{k=1}^{4} P(k)}. \tag{13}$$

In the example the results are as follows:

- $P(k = 1) = 0.5204$;

- $P(k = 2) = 0.2689$;
- $P(k = 3) = 0.1389$;
- $P(k = 4) = 0.0718$.

Note that the ratio between different $P(k)$ does not change.

Finally, we construct an adjusted probability distribution adding constants shown in second column of Table 6, as explained in Section 5. More specifically:

- $P_v(k = 1) = P(k = 1) + 0.13 = 0.6504$;
- $P_v(k = 2) = P(k = 2) - 0.10 = 0.1689$;
- $P_v(k = 3) = P(k = 3) - 0.03 = 0.1089$;
- $P_v(k = 4) = P(k = 4) = 0.0718$.

Note that the adjusted probability distribution is normalized, since:

$$\sum_{k=1}^{4} P_v(k) = \sum_{k=1}^{4} P(k) = 1. \tag{14}$$

$P_v(k)$ represents an empirically obtained probability distribution of $|Dst|$ level for a specific CME speed ($v = 2411$ km s^{-1}).

C. Probability distribution for CME/flare position distance from the centre of the solar disc (r), $P_r(k)$

The change of the $|Dst|$ distribution mean with CME/flare position distance from the centre of the solar disc, r, can be described with a power law function (see Figure 9):

$$m_{DST}(r) = a \cdot r^b. \tag{15}$$

Here $a = 30.95$, $b = -0.83$, and $|Dst|$ is expressed in nT. For a given $r = 0.4067$ Equation (15) gives distribution mean $m_{DST}(r) = 65.31$.

The geometric distribution mean is calculated using Equation (11). In the example $m_{GD} = 1.65$. The probability of the success in each trial, calculated using Equation (10) is $p = 0.60$.

Probabilities calculated using Equation (4) in Section 5, $P(k)$ are then:

- $P(k = 1) = 0.6049$;
- $P(k = 2) = 0.2390$;
- $P(k = 3) = 0.0944$;
- $P(k = 4) = 0.0373$.

The renormalized distribution, calculated using Equation (13) is:

- $P(k = 1) = 0.6200$;
- $P(k = 2) = 0.2450$;
- $P(k = 3) = 0.0968$;

- $P(k = 4) = 0.0382$.

Finally, the adjusted probability distribution (adding constants shown in third column of Table 6, as explained in Section 5) is:

- $P_r(k = 1) = P(k = 1) + 0.12 = 0.7400$;
- $P_r(k = 2) = P(k = 2) - 0.12 = 0.1250$;
- $P_r(k = 3) = P(k = 3) = 0.0968$;
- $P_r(k = 4) = P(k = 4) = 0.0382$.

$P_r(k)$ represents an empirically obtained probability distribution of $|Dst|$ level for a specific CME/flare position distance from the centre of the solar disc ($r = 0.4067$ solar radius).

D. Probability distribution for CME width (w), $P_w(k)$

The change of the $|Dst|$ distribution mean with CME width, w, can be described with a quadratic function (see Figure 9):

$$m_{DST}(w) = a \cdot w^2 + b \cdot w + c. \tag{16}$$

Here $a = 15.06$, $b = -34.60$, $c = 42.25$, and $|Dst|$ is expressed in nT. For a given $w = 3$ this gives distribution mean $m_{DST}(w) = 73.99$.

The geometric distribution mean is calculated using Equation (11). In the example $m_{GD} = 1.74$. The probability of the success in each trial, calculated using Equation (10) is $p = 0.57$.

Probabilities calculated using Equation (4) in Section 5, $P(k)$ are then:

- $P(k = 1) = 0.5747$;
- $P(k = 2) = 0.2444$;
- $P(k = 3) = 0.1039$;
- $P(k = 4) = 0.0442$.

The renormalized distribution, calculated using Equation (13) is:

- $P(k = 1) = 0.5942$;
- $P(k = 2) = 0.2527$;
- $P(k = 3) = 0.1075$;
- $P(k = 4) = 0.0457$.

Finally, the adjusted probability distribution (adding constants shown in fourth column of Table 6, as explained in Section 5) is:

- $P_w(k = 1) = P(k = 1) + 0.14 = 0.7342$;
- $P_w(k = 2) = P(k = 2) - 0.12 = 0.1327$;
- $P_w(k = 3) = P(k = 3) - 0.02 = 0.0875$;
- $P_w(k = 4) = P(k = 4) = 0.0457$.

$P_w(k)$ represents an empirically obtained probability distribution of $|Dst|$ level for a specific CME width ($w = 360°$, halo CME).

E. Probability distribution for flare class (f), $P_f(k)$

The change of the $|Dst|$ distribution mean with flare class, f, can be described with a quadratic function (see Figure 9):

$$m_{DST}(f) = a \cdot f^2 + b \cdot f + c. \tag{17}$$

Here $a = 10.41$, $b = -17.90$, $c = 46.93$, and $|Dst|$ is expressed in nT. For a given $f = 3$ this gives distribution mean $m_{DST}(f) = 86.92$.

The geometric distribution mean is calculated using Equation (11). In the example $m_{GD} = 1.87$. The probability of the success in each trial, calculated using Equation (10) is $p = 0.53$.

Probabilities calculated using Equation (4) in Section 5, $P(k)$ are then:

- $P(k = 1) = 0.5350$;
- $P(k = 2) = 0.2488$;
- $P(k = 3) = 0.1157$;
- $P(k = 4) = 0.0538$.

The renormalized distribution, calculated using Equation 13 is:

- $P(k = 1) = 0.5612$;
- $P(k = 2) = 0.2610$;
- $P(k = 3) = 0.1214$;
- $P(k = 4) = 0.0564$.

Finally, the adjusted probability distribution (adding constants shown in fifth column of Table 6, as explained in Section 5) is:

- $P_f(k = 1) = P(k = 1) + 0.15 = 0.7112$;
- $P_f(k = 2) = P(k = 2) - 0.12 = 0.1410$;
- $P_f(k = 3) = P(k = 3) - 0.02 = 0.1014$;
- $P_f(k = 4) = P(k = 4) - 0.01 = 0.0464$.

$P_f(k)$ represents an empirically obtained probability distribution of $|Dst|$ level for a specific flare class ($f = 3$, X class flare).

F. Probability distribution for interaction level (i), $P_i(k)$

The change of the $|Dst|$ distribution mean with interaction level, i, can be described with a power-law function (see Figure 9):

$$m_{DST}(i) = a \cdot i^b. \tag{18}$$

Here $a = 38.39$, $b = 0.49$, and $|Dst|$ is expressed in nT. For a given $i = 4$ this gives distribution mean $m_{DST}(i) = 65.77$.

The geometric distribution mean is calculated using Equation (11). In the example $m_{GD} = 1.66$. The probability of the success in each trial, calculated using Equation (10) is $p = 0.60$.

Probabilities calculated using Equation (4) in Section 5, $P(k)$ are then:

- $P(k = 1) = 0.5691$;
- $P(k = 2) = 0.2452$;
- $P(k = 3) = 0.1057$;
- $P(k = 4) = 0.0455$.

The renormalized distribution, calculated using Equation (13) is:

- $P(k = 1) = 0.5894$;
- $P(k = 2) = 0.2540$;
- $P(k = 3) = 0.1094$;
- $P(k = 4) = 0.0472$.

Finally, the adjusted probability distribution (adding constants shown in fifth column of Table 6, as explained in Section 5) is:

- $P_i(k = 1) = P(k = 1) + 0.15 = 0.7394$;
- $P_i(k = 2) = P(k = 2) - 0.13 = 0.1240$;
- $P_i(k = 3) = P(k = 3) - 0.01 = 0.0994$;
- $P_i(k = 4) = P(k = 4) - 0.01 = 0.0372$.

$P_i(k)$ represents an empirically obtained probability distribution of $|Dst|$ level for a specific interaction level ($i = 4$, interaction highly probable).

G. Combined probability distribution for set of key parameters (v, r, w, f, i), $P(|Dst|)$

Once we obtain the probability distribution of $|Dst|$ level for each of the key solar parameters (v, r, w, f, and i), their combined probability $P(k) = P(|Dst|)$ is calculated using Equation (8) in Section 5. For our example this gives:

- $P(k = 1) = P(|Dst| < 100nT) = 0.9982$;
- $P(k = 2) = P(100nT < |Dst| < 200nT) = 0.5253$;
- $P(k = 3) = P(200nT < |Dst| < 300nT) = 0.4056$;
- $P(k = 4) = P(|Dst| > 300nT) = 0.2178$.

Due to the fact that the set of parameters v, r, w, f, and i is not a complete set of independent variables for this distribution, this distribution is not normalized, i.e. $\sum P(|Dst|) \neq 1$. Therefore, it is necessary to renormalize the distribution (similarly to Equation (13)):

- $P(k = 1) = P(|Dst| < 100nT) = 0.4649$;
- $P(k = 2) = P(100nT < |Dst| < 200nT) = 0.2447$;
- $P(k = 3) = P(200nT < |Dst| < 300nT) = 0.1889$;
- $P(k = 4) = P(|Dst| > 300nT) = 0.1014$.

Note that the ratio between different $P(|Dst|)$ does not change. $P(|Dst|)$ represents an empirically obtained probability distribution of $|Dst|$ level for a specific set of key parameters ($v = 2411$, $r = 0.4067$, $w = 3$, $f = 3$, $i = 4$).

Acknowledgements The presented work has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n^{o} 263252 [COMESEP]. This work has been supported in part by Croatian Science Foundation under the project 6212 "Solar and Stellar Variability". This research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office (IAP P7/08 CHARM). L. Rodriguez acknowledges support from the Belgian Federal Science Policy Office through the ESA - PRODEX program. We are grateful to the SOHO LASCO CME catalog team for providing the CME data. This CME catalog is generated and maintained at the CDAW Data Center by NASA and The Catholic University of America in cooperation with the Naval Research Laboratory. SOHO is a project of international cooperation between ESA and NASA. We are also grateful to the Solar-Terrestrial Physics (STP) Division of NOAA's (National Oceanic and Atmospheric Administration) National Geophysical Data Center (NGDC) for providing solar flare data.

References

Akasofu, S.-I.: 1981, Energy coupling between the solar wind and the magnetosphere. *Space Sci. Rev.* **28**, 121 – 190. doi:10.1007/BF00218810.

Cargill, P.J.: 2004, On the Aerodynamic Drag Force Acting on Interplanetary Coronal Mass Ejections. *Solar Phys.* **221**, 135 – 149. doi:10.1023/B:SOLA.0000033366.10725.a2.

Cid, C., Cremades, H., Aran, A., Mandrini, C., Sanahuja, B., Schmieder, B., Menvielle, M., Rodriguez, L., Saiz, E., Cerrato, Y., Dasso, S., Jacobs, C., Lathuillere, C., Zhukov, A.: 2012, Can a halo CME from the limb be geoeffective? *J. Geophys. Res.* **117**, 11102. doi:10.1029/2012JA017536.

Dungey, J.W.: 1961, Interplanetary magnetic field and the auroral zones. *Phys. Rev. Lett.* **6**, 47 – 48. doi:10.1103/PhysRevLett.6.47. http://link.aps.org/doi/10.1103/PhysRevLett.6.47.

Farrugia, C., Berdichevsky, D.: 2004, Evolutionary signatures in complex ejecta and their driven shocks. *Ann. Geophys.* **22**, 3679 – 3698. doi:10.5194/angeo-22-3679-2004.

Gopalswamy, N., Yashiro, S., Akiyama, S.: 2007, Geoeffectiveness of halo coronal mass ejections. *J. Geophys. Res.* **112**, 6112. doi:10.1029/2006JA012149.

Gopalswamy, N., Makela, P., Yashiro, S., Davila, J.M.: 2012, The Relationship Between the Expansion Speed and Radial Speed of CMEs Confirmed Using Quadrature Observations of the 2011 February 15 CME. *Sun and Geosphere* **7**, 7 – 11.

Gosling, J.T., Bame, S.J., McComas, D.J., Phillips, J.L.: 1990, Coronal mass ejections and large geomagnetic storms. *Geophys. Res. Lett.* **17**, 901 – 904. doi:10.1029/GL017i007p00901.

Huttunen, K.E.J., Schwenn, R., Bothmer, V., Koskinen, H.E.J.: 2005, Properties and geoeffectiveness of magnetic clouds in the rising, maximum and early declining phases of solar cycle 23. *Ann. Geophys.* **23**, 625 – 641. doi:10.5194/angeo-23-625-2005.

Kim, R.-S., Cho, K.-S., Moon, Y.-J., Dryer, M., Lee, J., Yi, Y., Kim, K.-H., Wang, H., Park, Y.-D., Kim, Y.H.: 2010, An empirical model for prediction of geomagnetic storms using initially observed CME parameters at the Sun. *J. Geophys. Res.* **115**, 12108. doi:10.1029/2010JA015322.

Koskinen, H.E.J., Huttunen, K.E.J.: 2006, Geoeffectivity of Coronal Mass Ejections. *Space Sci. Rev.* **124**, 169 – 181. doi:10.1007/s11214-006-9103-0.

Lepping, R.P., Acũna, M.H., Burlaga, L.F., Farrell, W.M., Slavin, J.A., Schatten, K.H., Mariani, F., Ness, N.F., Neubauer, F.M., Whang, Y.C., Byrnes, J.B., Kennon, R.S., Panetta, P.V., Scheifele, J., Worley, E.M.: 1995, The Wind Magnetic Field Investigation. *Space Sci. Rev.* **71**, 207 – 229. doi:10.1007/BF00751330.

Maričić, D., Vršnak, B., Stanger, A.L., Veronig, A.M., Temmer, M., Roša, D.: 2007, Acceleration Phase of Coronal Mass Ejections: II. Synchronization of the Energy Release in the Associated Flare. *Solar Phys.* **241**, 99 – 112. doi:10.1007/s11207-007-0291-x.

McComas, D.J., Bame, S.J., Barker, P., Feldman, W.C., Phillips, J.L., Riley, P., Griffee, J.W.: 1998, Solar Wind Electron Proton Alpha Monitor (SWEPAM) for the Advanced Composition Explorer. *Space Sci. Rev.* **86**, 563 – 612. doi:10.1023/A:1005040232597.

Mishra, W., Srivastava, N., Chakrabarty, D.: 2014, Evolution and Consequences of Interacting CMEs of 2012 November 9-10 using STEREO/SECCHI and In Situ Observations. *Solar Physics-In press.*

Moon, Y.-J., Choe, G.S., Wang, H., Park, Y.D., Gopalswamy, N., Yang, G., Yashiro, S.: 2002, A Statistical Study of Two Classes of Coronal Mass Ejections. *Astrophys. J.* **581**, 694 – 702. doi:10.1086/344088.

Moon, Y.-J., Choe, G.S., Wang, H., Park, Y.D., Cheng, C.Z.: 2003, Relationship Between CME Kinematics and Flare Strength. *J. Korean Astron. Soc.* **36**, 61 – 66.

Möstl, C., Davies, J.A.: 2013, Speeds and Arrival Times of Solar Transients Approximated by Self-similar Expanding Circular Fronts. *Solar Phys.* **285**, 411 – 423. doi:10.1007/s11207-012-9978-8.

Möstl, C., Farrugia, C.J., Kilpua, E.K.J., Jian, L.K., Liu, Y., Eastwood, J.P., Harrison, R.A., Webb, D.F., Temmer, M., Odstrcil, D., Davies, J.A., Rollett, T., Luhmann, J.G., Nitta, N., Mulligan, T., Jensen, E.A., Forsyth, R., Lavraud, B., de Koning, C.A., Veronig, A.M., Galvin, A.B., Zhang, T.L., Anderson, B.J.: 2012, Multi-point Shock and Flux Rope Analysis of Multiple Interplanetary Coronal Mass Ejections around 2010 August 1 in the Inner Heliosphere. *Astrophys. J.* **758**, 10. doi:10.1088/0004-637X/758/1/10.

Ogilvie, K.W., Chornay, D.J., Fritzenreiter, R.J., Hunsaker, F., Keller, J., Lobell, J., Miller, G., Scudder, J.D., Sittler, E.C. Jr., Torbert, R.B., Bodet, D., Needell, G., Lazarus, A.J., Steinberg, J.T., Tappan, J.H., Mavretic, A., Gergin, E.: 1995, SWE, A Comprehensive Plasma Instrument for the Wind Spacecraft. *Space Sci. Rev.* **71**, 55 – 77. doi:10.1007/BF00751326.

Pitman, J.: 1993, *Probability*, Springer, New York.

Richardson, I.G., Cane, H.V.: 2010, Near-Earth Interplanetary Coronal Mass Ejections During Solar Cycle 23 (1996 - 2009): Catalog and Summary of Properties. *Solar Phys.* **264**, 189 – 237. doi:10.1007/s11207-010-9568-6.

Richardson, I.G., Cane, H.V.: 2011, Geoeffectiveness (Dst and Kp) of interplanetary coronal mass ejections during 1995-2009 and implications for storm forecasting. *Space Weather* **9**, 7005. doi:10.1029/2011SW000670.

Richardson, I.G., Webb, D.F., Zhang, J., Berdichevsky, D.B., Biesecker, D.A., Kasper, J.C., Kataoka, R., Steinberg, J.T., Thompson, B.J., Wu, C.-C., Zhukov, A.N.: 2006, Major geomagnetic storms (Dst $\leq$ -100 nT) generated by corotating interaction regions. *J. Geophys. Res.* **111**, 7. doi:10.1029/2005JA011476.

Rodriguez, L., Zhukov, A.N., Cid, C., Cerrato, Y., Saiz, E., Cremades, H., Dasso, S., Menvielle, M., Aran, A., Mandrini, C., Poedts, S., Schmieder, B.: 2009, Three frontside full halo coronal mass ejections with a nontypical geomagnetic response. *Space Weather* **7**, 6003. doi:10.1029/2008SW000453.

Russell, C.T., McPherron, R.L., Burton, R.K.: 1974, On the cause of geomagnetic storms. *J. Geophys. Res.* **79**, 1105. doi:10.1029/JA079i007p01105.

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

Smith, C.W., L'Heureux, J., Ness, N.F., Acuña, M.H., Burlaga, L.F., Scheifele, J.: 1998, The ACE Magnetic Fields Experiment. *Space Sci. Rev.* **86**, 613 – 632. doi:10.1023/A:1005092216668.

Srivastava, N.: 2005, A logistic regression model for predicting the occurrence of intense geomagnetic storms. *Ann. Geophys.* **23**(9), 2969 – 2974. doi:10.5194/angeo-23-2969-2005. http://www.ann-geophys.net/23/2969/2005/.

Srivastava, N.: 2006, The Challenge of Predicting the Occurrence of Intense Storms. *J. Astrophys. Astron.* **27**, 237 – 242. doi:10.1007/BF02702526.

Srivastava, N., Venkatakrishnan, P.: 2004, Solar and interplanetary sources of major geomagnetic storms during 1996-2002. *J. Geophys. Res.* **109**, 10103. doi:10.1029/2003JA010175.

Stirzaker, D.: 2003, *Elementary Probability*, Cambridge University Press, New York.

Stone, E.C., Frandsen, A.M., Mewaldt, R.A., Christian, E.R., Margolies, D., Ormes, J.F., Snow, F.: 1998, The Advanced Composition Explorer. *Space Sci. Rev.* **86**, 1 – 22. doi:10.1023/A:1005082526237.

Uwamahoro, J., McKinnell, L.A., Habarulema, J.B.: 2012, Estimating the geoeffectiveness of halo CMEs from associated solar and IP parameters using neural networks. *Ann. Geophys.* **30**, 963 – 972. doi:10.5194/angeo-30-963-2012.

Valach, F., Revallo, M., Bochníček, J., Hejda, P.: 2009, Solar energetic particle flux enhancement as a predictor of geomagnetic activity in a neural network-based model. *Space Weather* **7**, 4004. doi:10.1029/2008SW000421.

Verbanac, G., Mandea, M., Vršnak, B., Sentic, S.: 2011, Evolution of Solar and Geomagnetic Activity Indices, and Their Relationship: 1960 - 2001. *Solar Phys.* **271**, 183 – 195. doi:10.1007/s11207-011-9801-y.

Verbanac, G., Živković, S., Vršnak, B., Bandić, M., Hojsak, T.: 2013, Comparison of geoeffectiveness of coronal mass ejections and corotating interaction regions. *Astron. Astrophys.* **558**, A85. doi:10.1051/0004-6361/201220417.

Vršnak, B., Sudar, D., Ruždjak, D.: 2005, The CME-flare relationship: Are there really two types of CMEs? *Astron. Astrophys.* **435**, 1149 – 1157. doi:10.1051/0004-6361:20042166.

Vršnak, B., Ruždjak, D., Sudar, D., Gopalswamy, N.: 2004, Kinematics of coronal mass ejections between 2 and 30 solar radii. What can be learned about forces governing the eruption? *Astron. Astrophys.* **423**, 717 – 728. doi:10.1051/0004-6361:20047169.

Vršnak, B., Žic, T., Vrbanec, D., Temmer, M., Rollett, T., Möstl, C., Veronig, A., Čalogović, J., Dumbović, M., Lulić, S., Moon, Y.-J., Shanmugaraju, A.: 2013, Propagation of Interplanetary Coronal Mass Ejections: The Drag-Based Model. *Solar Phys.* **285**, 295 – 315. doi:10.1007/s11207-012-0035-4.

Yashiro, S., Gopalswamy, N., Michalek, G., St. Cyr, O.C., Plunkett, S.P., Rich, N.B., Howard, R.A.: 2004, A catalog of white light coronal mass ejections observed by the SOHO spacecraft. *J. Geophys. Res.* **109**, 7105. doi:10.1029/2003JA010282.

Yermolaev, Y.I., Nikolaeva, N.S., Lodkina, I.G., Yermolaev, M.Y.: 2012, Geoeffectiveness and efficiency of CIR, sheath, and ICME in generation of magnetic storms. *J. Geophys. Res.* **117**, 0. doi:10.1029/2011JA017139.

Zhang, J., Dere, K.P., Howard, R.A., Bothmer, V.: 2003, Identification of Solar Sources of Major Geomagnetic Storms between 1996 and 2000. *Astrophys. J.* **582**, 520 – 533. doi:10.1086/344611.

Zhang, J., Richardson, I.G., Webb, D.F., Gopalswamy, N., Huttunen, E., Kasper, J.C., Nitta, N.V., Poomvises, W., Thompson, B.J., Wu, C.-C., Yashiro, S., Zhukov, A.N.: 2007, Solar and interplanetary sources of major geomagnetic storms (Dst $\leq$ -100 nT) during 1996-2005. *J. Geophys. Res.* **112**, 10102. doi:10.1029/2007JA012321.