Distribution and clustering of fast coronal mass ejections

A. Ruzmaikin,[1] J. Feynman,[1] and S. A. Stoev[2]

Received 27 October 2010; revised 2 February 2011; accepted 10 February 2011; published 19 April 2011.

[1] The purpose of this paper is to investigate the statistical properties of high-speed coronal mass ejections (fast CMEs), which play a major role in Space Weather. We study the cumulative distribution of the initial CME speeds applying a new, advanced statistical method based on the scaling properties of averages of maximal speeds selected in time intervals of fixed sizes. This method allows us for the first time to obtain a systematic statistical description of the fast CME speeds. Using this method, we identify a self-similar (power law) high-speed portion of the spectrum of the speed maxima in the range of speeds from about 700 km/s to 2000 km/s. This self-similar range of the speed distribution provides a meaningful definition of "the fast" CMEs and indicates that these CMEs are produced by a process that is the same across the range of scales. The investigation of the temporal behavior of the fast CME events indicates that the time intervals between fast CMEs are not independent, i.e., fast CMEs arrive in clusters. We characterize the fast CMEs clustering by the exponent θ called the extremal index, which is the inverse of the averaged number of CMEs per cluster. An independent correlation analysis of the tail of the CME distribution confirms and further quantifies the temporal dependence among the fast CME events. To illustrate the predictive capabilities of the method, we identify clusters in the time series of CMEs with speeds greater than 1000 km/s and calculate their statistical characteristics such as the size and duration of the clusters. The method used in this paper can be applied to many other extreme geophysical events.

Citation: Ruzmaikin, A., J. Feynman, and S. A. Stoev (2011), Distribution and clustering of fast coronal mass ejections, *J. Geophys. Res.*, *116*, A04220, doi:10.1029/2010JA016247.

1. Introduction

[2] The coronal mass ejections (CMEs) vary widely in their speeds. When viewed near the Sun, some are very slow (<10 km/s) and others have very high speeds, even exceeding 2000 km/s [*Kahler*, 1987]. Among all CMEs the most interesting in the context of Space Weather are the high-speed CMEs. These fast CMEs and the shocks they generate in the solar wind are directly responsible for major geomagnetic storms [*Hirshberg and Colburn*, 1969; *Tsurutani and Gonzalez*, 1997; *Gopalswamy*, 2008] and solar energetic particle (SEP) events [*Reames*, 1999; *Li et al.*, 2005], which present hazards for spacecraft design and operation, for science instrumentation and astronauts. The causes of these enormous differences in CME speeds have not been identified as yet, but the differences in speed are likely presaged by differences in the buildup phases of the CMEs [*Feynman*, 1997]. The CMEs are associated with active regions and disappearing filaments, which appear randomly on the surface of the Sun. The frequency of their occurrence is regulated by the solar cycle. Observations have shown that active regions have a tendency to cluster, i.e., new magnetic fluxes preferably emerge in the vicinity of old ones [*Gaizauskas et al.*, 1983; *Harvey and Zwaan*, 1993]. The clusters may last for as many as six solar rotations and there are indications that the fastest CMEs preferably originate from them [*Ruzmaikin and Feynman*, 1998]. For example, the active region AR 8210 observed in April–May 1998 produced six CMEs with speeds greater than 1000 km/s [*Thompson et al.*, 2000]. During the famous Halloween period October–November 2003 most of the 80 observed CMEs originated from three active regions [*Gopalswamy et al.*, 2005]. Over 30 CMEs had speeds between 1000 km/s and 2000 km/s, and 7 CMEs had speeds exceeding 2000 km/s resulting in intense geomagnetic storms and large SEP events. These *fast* CMEs were launched in association with two of the solar active region clusters [*Feynman and Ruzmaikin*, 2004].

[3] Understanding and forecasting of the most hazardous extreme events in many geophysical phenomena (floods and major earthquakes, solar energetic particles, etc) requires the knowledge of the tails of probability distribution functions. The distributions of the intensity of these events are not Gaussian and are characterized by the extended high-intensity (heavy) tails, which give information that can be used in probabilistic prediction and physical understanding of the underlying physical processes. In practice the form of these high-energy tails is often estimated using a fit of an

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA.
[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA.

0148-0227/11/2010JA016247

empirical distribution to some known distribution function, such as a lognormal. The quality of the fit is difficult to evaluate because of scarcity of extreme events and a lack of precise mathematical techniques to do so. The curve fitting does not use specific data properties; it depends on the selected fraction of data, adjustable parameters and the skill of the researcher. Thus, [*Yurchyshyn et al.*, 2005] studied the distribution of plane-of-sky speeds determined for 4,315 CMEs detected by the SOHO LASCO in 1998–2001 and found that the speed distribution is non-Gaussian and can be fitted with a lognormal distribution. However recent advances in statistical methods of analysis have made it possible to find the form of the tail using approaches based on the properties of the data rather than the skill of curve fitting.

[4] In this paper we apply one of the new methods based on the use of scaling properties of the data maxima [*Stoev et al.*, 2006] to Coronal Mass Ejection (CME) speeds. We also examine the temporal dependence of the fast CMEs and by using the statistical properties of data maxima show that the time intervals between CMEs have a tendency to cluster. The method introduces two exponents: one defines the tail of the distribution function (i.e., extremes) and the other characterizes the clustering of extremes in time. We complement the calculation and discussion of these exponents with a supporting study of time correlation of extreme events obtained by different methods.

[5] Section 2 below describes the data set we used. Section 3 introduces the method used to quantify the form of the extreme tail and clustering of extreme events. Section 4 presents two exponents that characterize the distribution of speeds and clustering of fast CMEs. In section 5 we investigate the onset times of the fast CMEs (as observed in the solar corona) using one of the exponents and additional information obtained from the data. Section 6 summaries the results with a brief discussion of their possible implications for causes and consequences of fast CMEs.

2. The Data Set

[6] The plane of the sky CME speed propagation through the solar corona is measured by coronographic techniques. Here we use data from the Large Angle and Spectrometric Coronagraph Experiment on board the Solar and Heliospheric Observatory (LASCO SOHO) given in the catalog developed in cooperation with the Naval Research Laboratory and the Solar Data Analysis Center at the Goddard Space Flight Center and at the Center for Solar Physics and Space Weather at the Catholic University of America [*Gopalswamy et al.*, 2009]. The entries begin in January 1996. The model studies based on the STEREO mission observation show that actual 3D speeds are well correlated with the speeds determined by the LASCO [*Thernisien et al.*, 2009]. This justifies statistical analyses of the speeds from the LASCO catalog, although specific estimates of, say, a mean speed or a threshold speed for fast CMEs below are expected to be lower compared with estimates that would be found using the values of real 3D speeds.

[7] The CMEs in the LASCO catalog are listed according to the time of their first appearance above the C2 occulting disk and hence are spaced unevenly in time. Since our method of data analysis requires evenly spaced records, we form an hourly spaced time frame and assign the CMEs to the hour of their first appearance. Almost all CMEs are used, with no averaging or binning. The hours with no CMEs are assigned a zero speed. In a few cases when there is more than one CME in the same hour we use the CME with the highest speed, which is well justified by our method (see below) based on the investigation of speed maxima. To avoid the obvious nonstationarity due to solar cycle dependence we limit the data set to the high-activity part of solar cycle 23 (from January 1999 to December 2006, resulting in 9,408 CMEs). The speed we use for this paper is given in the catalog as obtained by the second order polynomial fit to the time-height measurements during the CME propagation through the solar corona. Note that even though we study properties of extreme (fast) CMEs the data input to the method includes all CMEs without preselection of those with high speeds. In particular, the speed at which the tail of the distribution function begins is not preselected but identified by the method.

[8] The observed speeds and their partial distribution function are shown in Figure 1. The distribution function of the speed for the time interval selected above clearly has a non-Gaussian form, as emphasized earlier by *Yurchyshyn et al.* [2005], with the peak at 263 km/s and an extended high-speed tail. The mean speed is 472 km/s. About 18% (1,746) of the CMEs have speeds exceeding 700 km/s, 6% (about 600 CMEs) have speeds exceeding 1000 km/s and less than 0.5% CMEs have speeds exceeding 2000 km/s.

3. The Method

[9] Here we briefly describe the method of studying extreme events, which we use to quantify the shape of the tail of the CME speeds distribution and the clustering of the fast CME onset times.

[10] We use the Max-Spectrum method based on investigating of averages of data maxima taken in all time intervals of fixed sizes when the size is progressively increased [*Stoev et al.*, 2006; *Hamidieh et al.*, 2009], see below for more detailed description. The method does not involve a fit to an empirically determined distribution function. It employs the scaling properties of a specific variable of the data, the data maxima observed at different time scales. The scaling approach, which had originally been used in turbulence studies (recall the Kolmogorov's power law preceded by attempts to fit a distribution function to the velocity increments) and now in many other applications, allows a natural extension of scaling when new data becomes available. It also allows an interpolation of the behavior of the variable beyond the limits of a given data set if there is no indication of any preferred value that could break the scaling. Let us briefly describe the method in application to the time series of CME speeds.

[11] Consider the time series of length N of the CME speeds $V(i)$, where $1 \le i \le N$. For each time scale index j ($j = 1, 2, 3, \ldots, [\log_2 N]$), we form nonoverlapping time blocks of length 2^j i. e. we progressively double the time scale. At each *fixed* scale j we calculate the maximum of the speed within each block:

$$D(j,k) = \max_{1 \le i \le 2^j} V\left(2^j(k-1)+i\right),\ k = 1, 2, \ldots, b_j,$$

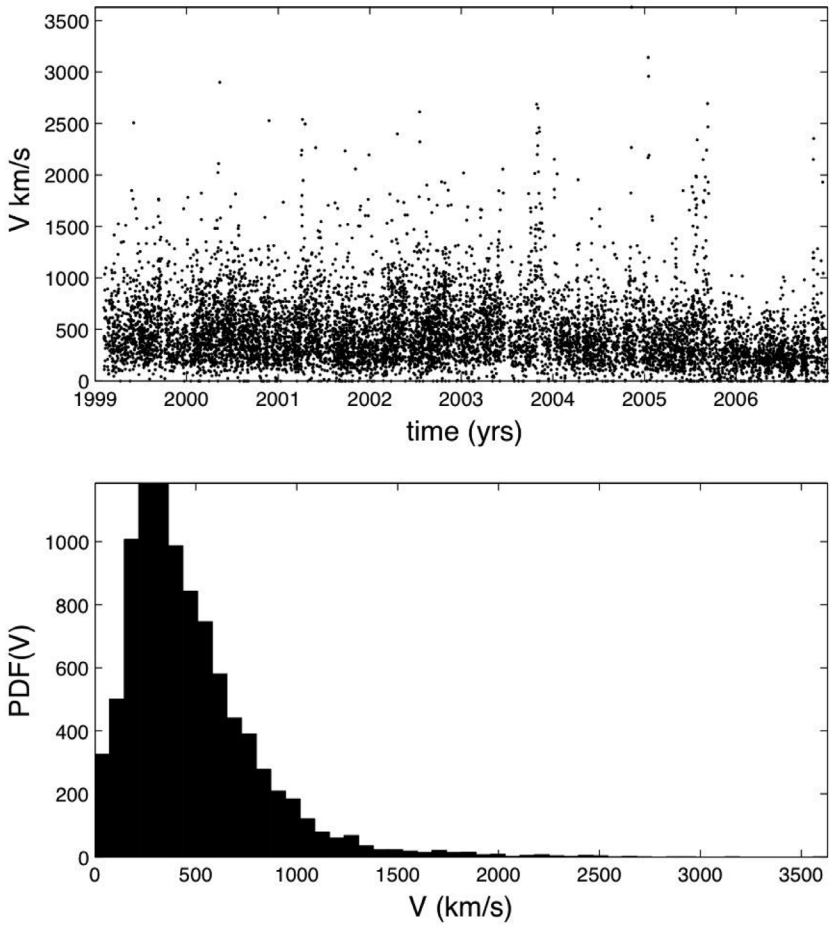

Figure 1. (top) The CME speeds taken in each hour from the LASCO catalog and positioned according to the hour of their occurrence. (bottom) Partial distribution function (PDF) of all CME speeds.

where $b_j = [N/2^j]$ is the number of blocks (of length 2^j) and i indexes the data points within the kth block. The index k defines the location of the block on the time axis. The log block size j plays the role of a time scale parameter. Observe that

$$D(j+1,k) = \max\{D(j,2k-1),\ D(j,2k)\},\ k = 1,2,\ldots,b_{j+1},$$

so the blocks of scale j are naturally nested in the blocks of scale $(j+1)$. Now, we average the logs of the block maxima $D(j, k)$ over all blocks at the fixed scale j:

$$Y(j) = \frac{1}{b_j}\sum_{k=1}^{b_j} \log_2 D(j,k).$$

The function $Y(j)$, i.e., a set of $[\log_2 N]$ numbers, is called the "Max-Spectrum" of the data. An important result, established by *Stoev et al.* [2006], is that if for a sufficiently large j

$$Y(j) \simeq j/\alpha + C, \tag{1}$$

where C is a constant and $\alpha > 0$, then the tail of the data distribution follows a power law with exponent α. If the tail were not of power law, say Exponential, Gaussian, or even Lognormal, the Max-Spectrum would level off at large scales.

[12] *Stoev et al.* [2006] proved that the exponent α is the same for both independent and dependent (correlated in time) data, provided that the time series are stationary and have the same distribution function. The dependence (related to the clustering of the times of extreme events) affects only the intercept in equation (1). That is, if we have *dependent* data with the same distribution function, then with the *same constant C*, equation (1) becomes

$$Y(j) \simeq j/\alpha + C + \log_2(\theta)/\alpha, \tag{2}$$

where the quantity θ $(0 < \theta \leq 1)$, is the extremal index introduced by *Leadbetter et al.* [1983].

[13] The extremal index is used in statistical studies to quantify the temporal clustering of the extreme events. Mathematical details that clarify the interpretation of the extremal index can be found by *Leadbetter et al.* [1983]. The notion of the extremal index is simple. The extremal index allows the distribution function of maxima of n *dependent* events to be presented as a distribution function of the maxima of roughly $n\theta$ *independent* events, i.e., to group the n-dependent events into $n\theta$-independent clusters. Thus the average number of events in a cluster is $1/\theta$. This

fact is used in section 5 below to identify clusters of fast CMEs.

[14] To be more formal, consider a stationary data time series X_t and let $M_n = \max_{1\le t\le n} X_t$ be the maximum of n consecutive data points. The time series X_t has an extremal index θ, if for large n the probability

$$\mathcal{P}\left(\frac{M_n - d_n}{c_n} \le x\right) = G^{\theta}(x), \text{ while } \mathcal{P}\left(\frac{M_n^* - d_n}{c_n} \le x\right) = G(x),$$

where M_n^* is the maximum of *n-independent* random variables X_t^* drawn from the same probability distribution as the original time series X_t, and $c_n > 0$, d_n are normalizing constants. The constants depend on the distribution function of the data. For example, if the X_t's are distributed as a power law with exponent $\alpha > 0$, then $c_n = n^{1/\alpha}$ and $d_n = 0$. It is important that the constants are the same whether the X_t's are *dependent* or *independent*. The $G(x)$ is the cumulative extreme value distribution:

$$G(x) = \exp\left\{-(1+\gamma(x-\mu)/\sigma)^{-1/\gamma}\right\},\ 1+\gamma(x-\mu)/\sigma > 0$$

and $G^{\theta}(x)$ is $G(x)$ to the power θ. The parameter $\sigma > 0$ plays the role of the scale, μ is the location and γ is the shape parameter of the distribution. If $\gamma \to 0$ we obtain the *Gumbel* cumulative distribution: $\exp\{-e^{-(x-\mu)/\sigma}\}$. If $\gamma < 0$, then the right-side tail is bounded and G becomes the *reversed Weibul* law. Finally, when $\gamma > 0$ the right-side tail decays like a power law and G is called the *Fréchet* distribution. We will show in section 4 that the *Fréchet* distribution models the fast CMEs.

[15] Note that the extremal index refers only to the temporal dependence between *extreme* events but not between all events. The smaller the index, the stronger is the extreme events interdependence that is exhibited by *clustering of time intervals between events*. In the limiting case $\theta = 1$ (independent events), consider the onset times t_i of CMEs with speeds exceeding a specified threshold speed U, which may be chosen as say 90th or 95th percentile of the speed distribution, or from physical considerations. Then the distribution of times between two consecutive onsets of CMEs $\tau_i = t_i - t_{i-1}$, $i = 1, 2, ..$ is simply $\mathcal{P}(\tau = k) = (1-p)^{k-1}p$, where $k = 1, 2, 3,\ldots$, marks the discretized time and $p = p(U)$ denotes the probability of occurrence of one CME in a unit of time. For large U, p is small and this distribution converges to an exponential distribution with the expectation value $1/p = 1/\mathcal{P}(V_t > U)$.

[16] Equations (1) and (2) suggest a method of estimating both α and θ [*Stoev et al.*, 2006; *Hamidieh et al.*, 2009]. The inverse exponent $1/\alpha$ is obtained as a slope of the line fitted to the Max-Spectrum of the data. The best linear fit outlines the self-similar part of the Max-Spectrum. We should take into account that in practice the larger the scale j, the fewer the block maxima $D(j, k)$ (indexed by k) and the greater the variability of the Max-Spectrum statistic $Y(j)$. The best way to deal with this problem is applying, as we do here, the method of generalized least squares, which accounts for the bias variance tradeoff [*Stoev et al.*, 2006].

[17] Taking into account equations (1) and (2), we can obtain estimates of the extremal index. By permuting the data with a substitute of data points (bootstrap) or by simply randomly permuting the original data time series, we obtain a time series V_i^*, $1 \le i \le N$, which has the same distribution function as the original data set but the dependence (i.e., correlations between data points) has been destroyed. Carrying this out we create a large set of pseudo time series in which the original data dependence is destroyed and the events may be viewed as nearly independent in time. For each such time series, we compute a Max-Spectrum, $Y^*(j)$, $1 \le j \le [\log_2 N]$ that satisfies equation (1). The Max-Spectrum of the original data $Y(j)$ satisfies equation (2) with the same constant C, thus the difference between two spectra yields an estimate of θ:

$$\widehat{\theta}(j) = 2^{-\widehat{\alpha}\left(Y^*(j) - Y(j)\right)}, \tag{3}$$

where $\widehat{\alpha}$ stands for an estimate of the tail exponent α, obtained from the slope of the Max-Spectrum. Since we have a large sample of pseudo-independent time series, we obtain many realizations of $\widehat{\theta}(j)$ at each scale j. The median or the mean of these estimates can be taken as a point estimator of θ at the scale j. The whole sample of estimates can be used to quantify the estimation error at each scale.

4. The Distribution and the Clustering of Fast CMEs

[18] Using the method described above we calculate the values of the exponent α and of the extremal index θ for the CME speeds. The resulting Max-Spectrum of the CME speeds is shown in Figure 2. Our best fit to the slope gives evidence that the cumulative distribution function of the CME speeds has a Fréchet type power law tail, with the exponent $\alpha = 3.4$. The lower boundary of the Max Spectrum identifies the onset of the power law tail, i.e., the corresponding speed threshold, and a self-similar range. This gives a meaningful definition of "the fast" CMEs. Specifically, we find that the Max Spectrum above 700 km/s is self-similar. Bear in mind the analogy with the standard, self-similar cascade process in turbulence, which is fully defined by a Kolmogorov-type spectral index, we conjecture that the physical process leading to the fast CMEs production is the same from about 700 km/s to the highest velocities in the data set.

[19] To be more confident, we made estimates of the extremal index shown in Figure 3 by two independent methods. The first estimate is obtained by the Max-Spectrum method (Figure 3, top) and the second (Figure 3, bottom) by an alternative estimator based on the use of the data percentiles quantifying the average number of CME's of speeds exceeding U that arrive in a "cluster" [*Ferro and Segers*, 2003]. The Ferro-Segers method produces a more stable index than the Max-Spectrum for small speed thresholds. This is because the maxima taken in blocks of data in the Max-Spectrum method include both extreme ($\ge U$ km/s) and nonextreme ($< U$ km/s) speeds. Hence the blocks of small sizes (small scales) are dominated by more numerous maxima having small speeds. The Max-Spectrum performs well when the size of blocks is larger, i.e., at sufficiently large scales. The Max-Spectrum method has the advantage of providing "confidence intervals" as illustrated with the histogram (Figure 4) plotted for the thresholds

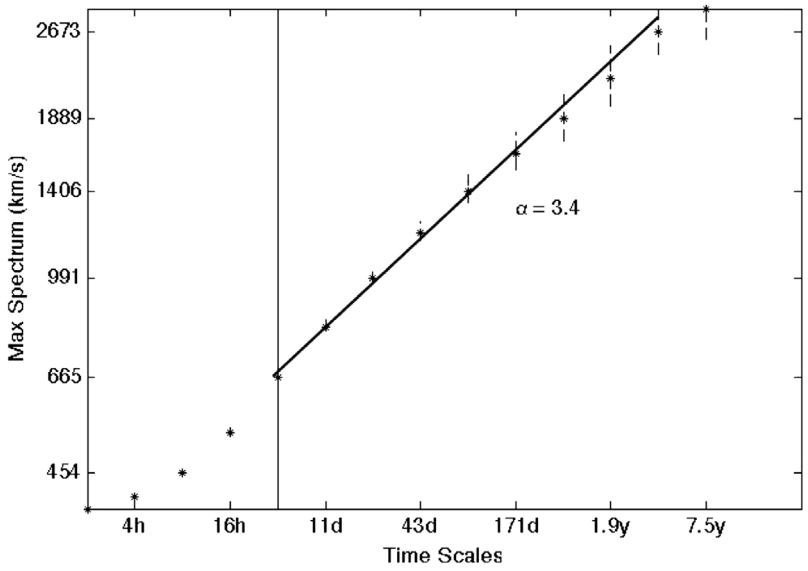

Figure 2. The Max-Spectrum of the CME speeds at progressively increased time scales. The error bars are estimated using the generalized regression [*Stoev et al.*, 2006], corresponding to 95% confidence intervals. We converted the $\log_2$ units for Y(j) into km/s and converted the scales j into time units 2^j. The vertical line segment indicates the starting scale selected for the evaluation of α. The speed at this scale may be interpreted as the beginning of the distribution function tail thus defining the fast CMEs.

1000–2300 km/s. The resulting empirical 95% confidence interval for θ is from 0.33 to 0.60 with a midpoint $\theta = 0.49 \approx 0.5$. The values 0.4–0.6 seen on the bottom for the Ferro-Segers estimator also fall in this interval. The rapidly growing statistical error for the "most extreme" CMEs suggests that neither of the two methods should be used for speeds above 2300 km/s due to insufficient number of available data points. The θ in the range 0.3–0.6 with the mean 0.5 can be taken as an estimate of the extremal index.

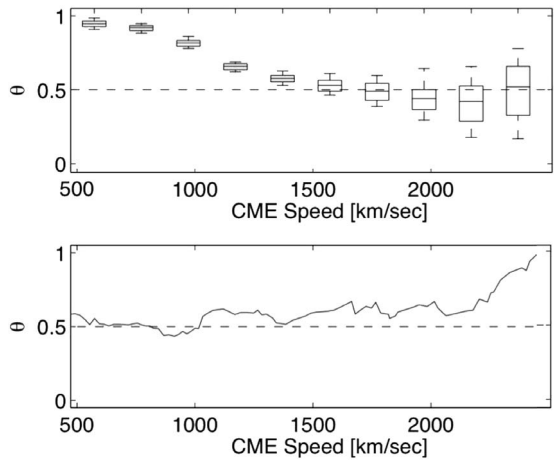

Figure 3. (top) The extremal index obtained by the Max-Spectrum method. The box plots for each time scale are obtained from 100 independent realizations of the randomized $\widehat{\theta}$, as explained in the text. The central mark in a box is the median; the box edges are the 10th and 90th percentiles and whiskers extend to the most extreme data points. (bottom) The extremal index obtained by the Segers-Ferro method.

The inverse value of the index gives an estimate of an average cluster size 2–3, i.e., *on average* an appearance of a fast CME will be followed by one or two other fast CMEs. These estimates can be further justified by the asymptotic statistical theory of *Hsing et al.* [1988].

[20] Although the extremal index provides statistical evidence for significant temporal dependence between the fast CMEs it is also imperative to investigate the correlation of the fast CMEs. The traditional correlation function is not useful for this purpose because we consider only extreme speed CMEs. Instead we consider the probability λ_k of observing another fast CME k time lags after a fast CME has already been observed, i.e.,

$$\lambda_k(U) = P(V_k > U | V_0 > U), \; k = 1, 2, \ldots \tag{4}$$

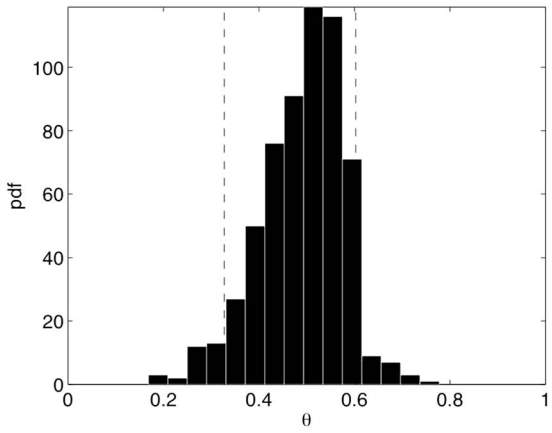

Figure 4. The histogram of the $\widehat{\theta}$ for speed thresholds from 1000 to 2300 km/s.

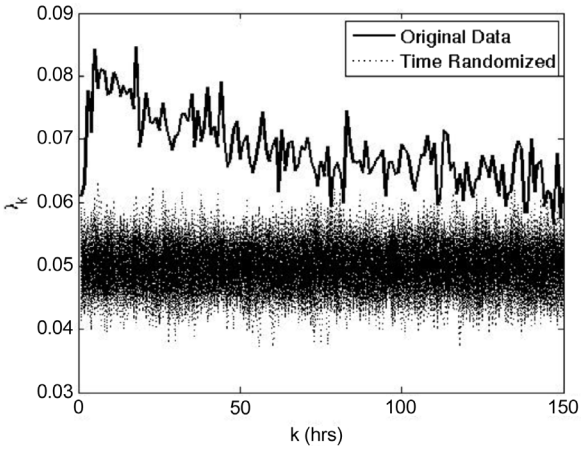

Figure 5. Tail correlation parameter λ_k for the fast CMEs ($>U$ = 700 km/s) as a function of the time lag given by the solid line. The cloud of dots below indicates 100 estimates of this parameter for randomly rearranged CMEs. The mean level(5%) of this cloud is well below the parameter calculated from the data.

for a threshold speed U. Now, if the V_k's were time independent, this conditional probability would equal the unconditional probability $P(V_k > U)$. Therefore, the V_k's are statistically dependent in time if λ_k is significantly different from $P(V_k > U)$. The parameter λ_k can be estimated with the following empirical statistic:

$$\widehat{\lambda}_k = \left(\frac{\sum_{j=1}^{n-k} I(V_{j+k} > U, V_j > U)}{n-k}\right) \Big/ \left(\frac{\sum_{j=1}^{n} I(V_j > U)}{n}\right), \quad (5)$$

where $I(A)$ equals one if the event A occurs and zero otherwise.

[21] Figure 5 shows the estimate of λ_k as a function of the lag k for the CME data set with the threshold speed U corresponding to the 95th percentile of the data (1000 km/s). To test the statistical significance of λ_k we randomized the order of the V_k's and calculated λ for these randomized (independent) in time data. This calculation was repeated independently 100 times and the resulting λ_k's are shown by the dots. One can see that λ is significantly larger than the unconditional probability $P(V_0 > U)$ (equal to 5% in this case) for the lags less than at least 150 h thus confirming our conclusion that the fast CMEs are temporally dependent. This empirical estimate of λ_k may be used to predict the likelihood of a CME as fast or faster than U km/s k hours in the future.

5. Size and Duration of CME Clusters

[22] It is useful to demonstrate the interdependence of fast CME onset times directly by using the observed time intervals between the CMEs. In Figure 6 we compare the distribution of the observed time intervals between events

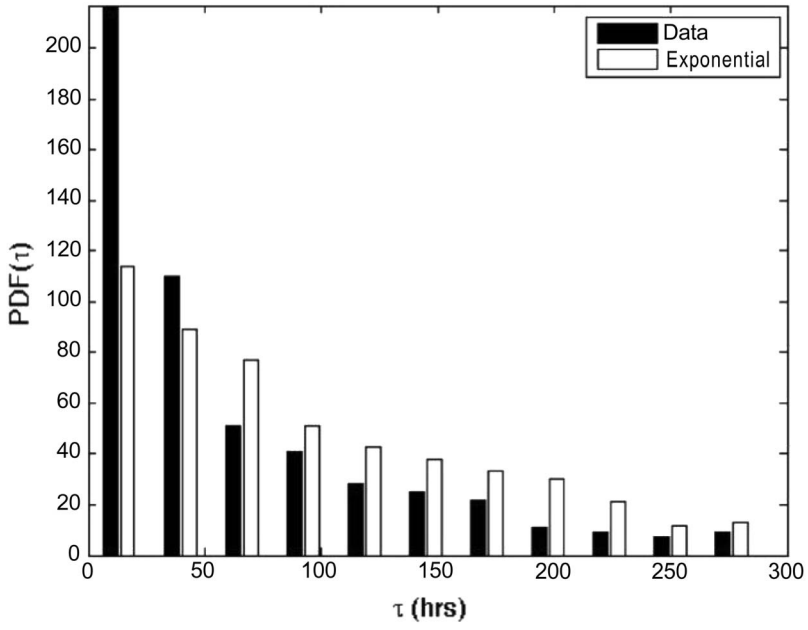

Figure 6. Distribution function of time intervals between successive CMEs with speeds exceeding 1000 km/s (black) versus the Exponential distribution function of time intervals for a randomized data (white). The sample, with which the distributions are built, consists of 586 events. The Exponential distribution has been generated in MATLAB by the operator log(1./rand(1,n)), where rand(1,n) are uniform (0,1) random numbers of length n, and then normalized using the mean value of the data time intervals. Both distributions have the same standard deviation, which is also a mean value for the randomized distribution. The peak near zero for the real CME data indicates the dominance of small time intervals compared to the times from the random distribution (i.e., clustering of extremes).

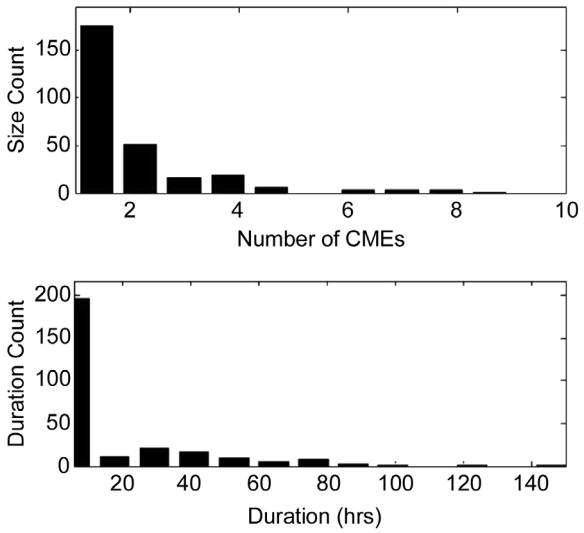

Figure 7. Distribution of cluster sizes and duration of the CME clusters for fast CMEs with speeds exceeding 1000 km/s.

for CMEs with speeds > 1000 km/s with the Exponential distribution expected if these fast CMEs occurred independently of one another. Figure 6 clearly shows that the Exponential time distribution does not fit our data set. There are about twice as many time intervals of duration less than about a day than would be expected for independent events. This means that many fast CMEs essentially arrive in groups that are closely spaced together (clusters). The distribution of time intervals between two clusters of fast CMEs follows the Exponential distribution.

[23] The mean time interval between CMEs within a cluster depends on the speed threshold. It is known from theory [*Hsing et al.*, 1988] that if one focuses on asymptotically larger and larger thresholds U, the time intervals τ_i between the fast CMEs will converge (under time-rescaling) to a cluster Poisson process, which is similar to a Poisson process, but with several (random number) of events arriving clustered in time. The average number of events in a cluster defined by a given threshold is $1/\theta$.

[24] To obtain more detailed information about the clusters it is instructive to apply the statistical methodology, called "de-clustering", which employs the extremal index [*Ferro and Segers*, 2003].

[25] First, the extremal index allows us to estimate the number of clusters. Indeed, if our time series consists of n extreme events (i.e., n CMEs with speeds exceeding a threshold U), they are on average grouped into $\theta \times n$ clusters. Second, there is a useful concept of a "de-clustering threshold time" τ_c. Consider the time intervals τ_i between consecutive fast CMEs. If the time interval between two fast CMEs is less than τ_c, then these CMEs can be grouped into a cluster. To determine the "de-clustering threshold time," which separates intracluster time intervals from intercluster time intervals, we consider the sorted collection of all times between consecutive fast CMEs

$$\tau_1 \le \tau_2 \le \cdots \le \tau_c \le \ldots \le \tau_{n-1} \tag{6}$$

and take τ_c as the $\theta \times n$th largest among them [*Ferro and Segers*, 2003]. As an example consider a threshold U = 1000 km/s and $\theta = 0.5$. With this threshold we have n = 586 fast CMEs with the "de-clustering time" τ_c = 42 h. (A close de-clustering time can be obtained by comparison the distribution of observed time intervals with the Exponential distribution as has been shown in Figure 6.) By identifying the start and end times of intervals exceeding τ_c we can count the number of clusters with 2, 3, .. and more members and the duration of these clusters. The average duration time within a cluster is 18 h with standard error 2 h and the distribution of the cluster durations is highly skewed (see Figure 7 (bottom)).

[26] Table 1 provides more detailed information about the probability and the corresponding duration of clusters as a function of their size (number of CMEs in the cluster). We see that about 30% of the fast CMEs are single. The rest of the fast CMEs are in clusters of different sizes. There is a statistically significant proportion (about 35%) of clusters with five or more members, which have an average duration of about 110 h. This duration is in agreement with the estimate shown in Figure 5.

[27] These empirical findings confirm and quantify the presence of temporal dependence of the CMEs. Similar estimates can be made using different threshold speeds. Figure 8 shows estimates of how the ratio of the cluster

Table 1. Example of Predictive Statistics for the Clusters of CMEs With Speed Exceeding 1000 km/s[a]

Size	Number of Clusters	Number of CMEs in Clusters	Recording Probabilities	Mean Duration (h)
1	177	177	0.61 (0.03)	-
2	53	106	0.18 (0.02)	20.1 (1.7)
3	18	54	0.06 (0.01)	39.7 (3.8)
4	20	80	0.07 (0.01)	56.8 (4.5)
5	7	35	0.02 (0.01)	70 (7.2)
>5	17	169	0.06 (0.01)	107.7 (10.6)

[a]The first column (size) lists the number of CMEs in the cluster. The second and third columns give the number of clusters of this size and total number of CMEs in these clusters. The fourth column provides estimates and standard error (in parentheses) of the probabilities that a cluster of the corresponding size is recorded. The last column lists the expected mean durations of the clusters (with standard error in parentheses).

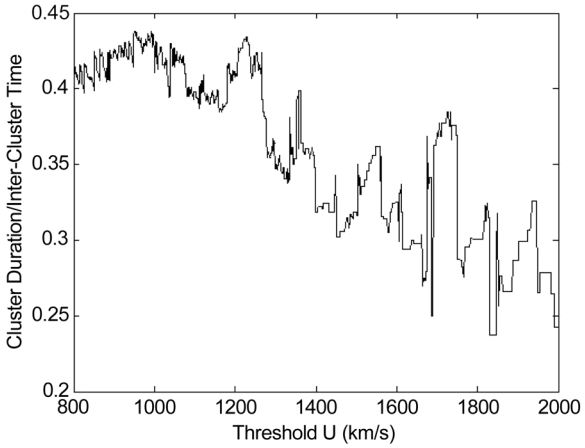

Figure 8. The ratio of cluster duration to intercluster time as a function of the speed threshold.

duration to the time between the clusters depends on the speed threshold.

6. Conclusions

[28] The Max-Spectrum method has allowed us for the first time to obtain a systematic statistical description of the fast CME speeds. We find by analyzing the data maxima that the fast CME speeds have a Fréchet type self-similar distribution, i.e., asymptotically follow a power law with the power exponent 3.4 in the range of speeds from about 700 km/s to 2000 km/s. Our tests show that the value of the exponent weakly depends on the size of the data sample thus evidencing on stationary of the Max-Spectrum. For example after splitting the studied time series into two parts we find $\alpha = 3.1$ for the first half and $\alpha = 3.2$ for the second half of the time series.

[29] In statistics the power law tails are called "heavy". They are commonly observed in financial and Internet traffic data. The fact that the CME speed has a heavy (power law) tail means that the fast ones are produced with much larger probability than one would expect from the standard normal or exponential distribution. The lognormal distribution does not belong to the class of heavy-tailed distributions. Practically, for any large but finite data sample the lognormal distribution, which has two adjustable parameters (mean and standard deviation), approximates a power law distribution rather well. And this is probably the reason why the lognormal distribution is often used. However, in contrast with the heavy-tailed distributions, the lognormal approximation needs reevaluation of the fitting parameters as new data become available.

[30] The finding of the self-similar range of the speed distribution provides a meaningful definition of "the fast" CMEs and has an important consequence for our understanding of the physical process responsible for their generation. As indicated by observations [*Feynman*, 1997], the fast CMEs apparently originate from the clusters of emerging magnetic flux on the solar surface. These powerful agglomerates of activity produce more energetic (fast) CMEs compared with single active regions. Thus the clustering of active regions apparently simulates the clustering of the fast CMEs. The existence of a self-similar range of scales and related speeds means that the physical process responsible for producing the series of fast CMEs is the same over the ranged scales and differs in some way from the process that generates the slower CMEs. However the connection between active regions and fast CMEs is not straightforward already because not every active region produces a CME. Active regions are generated by solar dynamo process, and clustering of active regions is, least conceptually, understood in the context of the solar dynamo [*Ruzmaikin*, 1998]. But we do not know yet how and which cluster of solar activity produces a multiple set of fast CMEs.

[31] Another finding from our study has a potential for developing a capability for statistical prediction of fast CMEs. We find that the onset times of the fast CMEs are not independent, as would be expected according to a standard Poisson process, but they tend to cluster. This "clustering phenomenon" is described in the context of Extreme Value Theory by the extremal index [*Leadbetter et al.*, 1983]. Using the extremal index we estimated the critical time scale that separates the time intervals between clusters from time intervals between CMEs within a cluster as a function of the speed threshold. Note that both exponents, α and θ, can be useful in statistical forecasting of fast CME. The exponent α gives us a range of fast CMEs speed thresholds, and θ is used to separating the time intervals between the clusters from time intervals between CMEs within the clusters.

[32] In the Space Weather context clustering implies a serial impact of CMEs on interplanetary environments. For example, if a CME over 1000 km/s occurs one should expect with probability 60% another CME with that speed or faster within the next 2 days (see Table 1 for more details). Lowering the speed threshold leads to more fast CMEs per cluster and longer duration of the clusters relative to the times between clusters (Figure 8). The clustering in time of fast CMEs also means that the process (mechanism) of their production must include the correlation (memory) between the subsequently launched CMEs. In other words, the process should be not simply additive (this type of process leads to normal, Gaussian distribution) but multiplicative, similar to the spread of the forest fires or to cascade process in the turbulent inertial range.

[33] An important consequence of the CME clustering for the Space Weather has had in fact been used earlier in an empirical *definition of a SEP event* [*Feynman et al.*, 1993, 2002]. A high-flux SEP event (closely associated with the fast CMEs) was defined as a cluster of the fluxes and fluences appearing over several days. Thus defined SEP event typically involves many successive increases in particle flux. It has also been shown that the time between the SEP events is distributed according to the exponential law of the Poisson process, while the timing between all SEPs does not follow this distribution. (For a discussion of possible generalization to a time-dependent Poisson process see *Jiggens and Gabriel* [2009].) This definition of the SEP event is widely used in space environment models employed for the designs of space missions. The methods and results presented in our paper put a firm scientific basis for definitions of "extreme Space Weather events," including fast CME and extreme SEP events. Table 1 gives an example of statistical estimates that can be a useful guide in developing techniques for prediction of fast CMEs and related to them SEPs.

[34] **Acknowledgments.** We are grateful to anonymous reviewers for helpful critical comments. This work was supported in part by the Jet Propulsion Laboratory of the California Institute of Technology, under a contract with the National Aeronautics and Space Administration. S. Stoev was partially supported by the NSF grant DMS-0806094.

[35] Philippa Browning thanks Stephen B. Gabriel and another reviewer for their assistance in evaluating this manuscript.

References

Ferro, C. A. T., and J. Segers (2003), Inference for clusters of extremes, *J. R. Stat. Soc., Ser. B*, *65*, 545–556.

Feynman, J. (1997), Evolving magnetic structures and their relationship to the coronal mass ejections, in *Coronal Mass Ejections*, *Geophys. Monogr. Ser.*, vol. 99, edited by N. Crooker, J. A. Joselyn, and J. Feynman, pp. 49–56, AGU, Washington, D. C.

Feynman, J., and A. Ruzmaikin (2004), A high-speed erupting-prominence CME: A bridge between types, *Sol. Phys.*, *219*, 301–313.

Feynman, J., G. Spitale, J. Wang, and S. Gabriel (1993), Interplanetary proton fluence model: JPL 1991, *J. Geophys. Res.*, *98*, 13,281–13,294.

Feynman, J., A. Ruzmaikin, and V. Berdichevsky (2002), The JPL proton fluence model: An update, *J. Atmos. Sol. Terr. Phys.*, *64*, 1679–1686.

Gaizauskas, V., K. L. Harvey, J. W. Harvey, and C. Zwaan (1983), Large-scale patterns formed by solar active regions during the ascending phase of cycle 21, *Astrophys. J.*, *265*, 1056–1065.

Gopalswamy, N. (2008), Solar connections of geoeffective magnetic structures, *J. Atmos. Sol. Terr. Phys.*, *70*, 2078–2100.

Gopalswamy, N., S. Yashiro, Y. Liu, G. Michalek, A. Vourlidas, M. L. Kaiser, and R. A. Howard (2005), Coronal mass ejections and other extreme characteristics of the 2003 October–November solar eruptions, *J. Geophys. Res.*, *110*, A09S15, doi:10.1029/2004JA010958.

Gopalswamy, N., S. Yashiro, G. Michalek, G. Stenborg, A. Vourlidas, S. Freeland, and R. Howard (2009), The SOHO/LASCO CME Catalog, *Earth Moon Planets*, *104*, 295–313.

Hamidieh, K., S. Stoev, and G. Michailidis (2009), On the estimation of the extremal index based on scaling and resampling, *J. Comut. Graph. Stat.*, *18*, 731–755, doi:10.1198/cgs.2009.08065.

Harvey, K. L., and C. Zwaan (1993), Properties and emergence patterns of bipolar active regions, *Sol. Phys.*, *148*, 85–118.

Hirshberg, J., and D. S. Colburn (1969), Interplanetary field and geomagnetic variations—a unified view, *Planet. Space Sci.*, *17*, 1183–1206.

Hsing, T., J. Hüsler, and M. R. Leadbetter (1988), On the exceedance point process for a stationary sequence, *Probab. Theory Related Fields*, *78*, 97–112.

Jiggens, P. T. A., and S. B. Gabriel (2009), Time distributions of solar energetic particle events: Are SEPEs really random?, *J. Geophys. Res.*, *114*, A10105, doi:10.1029/2009JA014291.

Kahler, S. (1987), Coronal mass ejections, *Rev. Geophys.*, *25*, 663–675, doi:10.1029/RG025i003p00663.

Leadbetter, M. R., G. Lindgren, and H. Rootzen (1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.

Li, G., G. P. Zank, and W. K. M. Rice (2005), Acceleration and transport of heavy ions at coronal mass ejection-driven shocks, *J. Geophys. Res.*, *110*, A06104, doi:10.1029/2004JA010600.

Reames, D. V. (1999), Particle acceleration at the Sun and in the heliosphere, *Space Sci. Rev.*, *90*, 413–491.

Ruzmaikin, A. (1998), Clustering of emerging magnetic flux, *Sol. Phys.*, *181*, 1–12.

Ruzmaikin, A., and J. Feynman (1998), Fast CMEs and their association with clustering of emerging flux, in *Journal of Physics of Space Plasmas*, edited by T. Cheng and J. R. Jasperse, pp. 295–300, MIT Cent. for Theor. Phys., Cambridge, Mass.

Stoev, S. A., G. Michailidis, and M. S. Taqqu (2006), Estimating heavy-tail exponents through max self-similarity, *Tech. Rep. 445*, 447, Univ. of Mich., Ann Harbor.

Thernisien, A., A. Vourlidas, and R. A. Howard (2009), Forward modeling of coronal mass ejections using STEREO/SECCHI data, *Sol. Phys.*, *256*, 111–130, doi:10.1007/s11207-009-9346-5.

Thompson, B. E., E. W. Cliver, N. Nitta, C. Delannee, and J.-P. Delaboudiniere (2000), Coronal dimmings and energetic CMEs in April–May, 1998, *Geophys. Res. Lett.*, *27*, 1431–1435.

Tsurutani, B. T., and W. D. Gonzalez (1997), The interplanetary causes of magnetic storms: A review, in *Magnetic Storms*, *Geophys. Monogr. Ser.*, vol. 98, pp. 77–93, AGU, Washington, D. C.

Yurchyshyn, Y., S. Yashiro, V. Abramenko, H. Wang, and N. Gopalswamy (2005), Statistical distributions of speeds of coronal mass ejections, *Astrophys. J.*, *619*, 5099–6030.

J. Feynman and A. Ruzmaikin, Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA. (joan.feynman@jpl.nasa.gov; alexander.ruzmaikin@jpl.nasa.gov)

S. A. Stoev, Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, USA. (sstoev@umich.edu)