CHAPTER

STATISTICS OF EXTREME SPACE WEATHER EVENTS

5

Pete Riley
Predictive Science Inc., San Diego, CA, United States

CHAPTER OUTLINE

1 INTRODUCTION

Space weather refers to the state and future conditions of the space environment surrounding the Earth, including within the Earth's magnetosphere. As a discipline, it is predominantly operational in nature. Extreme space weather refers to those conditions that are so far removed from the norm that they are rare. Unfortunately, defining what makes a space weather event, or space weather conditions extreme, is difficult. An event, for example, can be extreme with respect to one parameter (say, the *Dst* index) but not another (say, coronal mass ejection (CME) speed). However, some events, such as those occurring on September 1, 1859 (e.g., Riley, 2012) and July 23, 2012 (e.g., Riley et al., 2016), are undoubtedly extreme.

Extreme Events in Geospace. https://doi.org/10.1016/B978-0-12-812700-1.00005-4

The Carrington event is perhaps the quintessential space weather event and has been studied in great detail, at least to the extent possible because of the limited data associated with it (e.g., Riley, 2012). Similarly, the July 23, 2012, event, which was observed by the STEREO A spacecraft, located at 1 AU from the Sun but away from the Sun-Earth line, has significantly more solar and heliospheric data but no geomagnetic data because the ICME did not hit the Earth (Russell et al., 2013; Baker et al., 2013; Liu et al., 2014). While there are a number of similarities between the two, there are also important distinctions. However, both can serve as proxies for a generic "extreme" event. And, while much can be learned from studying specific events, broad extrapolations may lead to inaccurate inferences. In this chapter we summarize our current knowledge with respect to the statistics of extreme space weather events, and, in particular, in estimating the probability of occurrence over the next decade.

A number of studies have considered the properties of extreme solar events. Koons (2001), for example, applied extreme value statistics to show that the extreme events observed within several space weather datasets, including the magnetic index *Sp*, are well fit by extreme value models. Tsubouchi and Omura (2007) assumed a power-law distribution for magnetic storms to infer that an event as large as the March 1989 storm ($|Dst| > 280$ nT) would occur every 60 years or so. Finally, it should be noted that quasipower-law behavior is not limited to space physics, but can be found in a wide range of natural hazards, such as earthquakes, volcanic eruptions, floods, wildfires, and landslides, to name just a few (Sachs et al., 2012).

Riley (2012) first considered the likelihood of a Carrington-like event occurring on a decadal timescale. They showed that even using simple "time-to-event" analysis, this likely was ~9%. Using more sophisticated statistical tools, they suggested that this number ranged from a few percent up to 12%, depending on the parameter under study (i.e., how one defined an "extreme" event). They considered a number of solar, heliospheric, and magnetospheric indices or parameters. Although solar flares (which have often been connected with power-law distributions) would seem to be an ideal parameter, Riley (2012) showed that this was a questionable inference, and that these data were not well suited for further analysis. Similarly, CME speed appeared to show a broken power-law distribution, with a knee just above 2000 km/s. The *Dst* index generally showed what was interpreted to be a power-law distribution, as did solar proton events (SPEs), as evidenced by >30 MeV proton fluences.

Love (2012) independently also considered the probability of rare geomagnetic storms, but focused on an estimate of the uncertainties associated with them, which were only indirectly addressed by Riley (2012). He found that the 95% confidence intervals for a storm exceeding −589 nT lay between 3.4% and 38.6%, that is, from quite unlikely to very probable. Thus, a crucial part of the analysis that we must attempt to refine is not so much the forecast estimate but a reduction in the confidence intervals.

Love et al. (2015) questioned the basic assumption that the data, and the *Dst* index in particular, could be adequately described by a power-law distribution. By including more moderate storms in their analysis, they argued that a log-normal distribution was a better fit to the data than a pure power-law distribution. Surprisingly, this difference only modified the likelihood modestly.

Most recently, Riley and Love (2017) have incorporated and generalized the concepts from the previous studies to produce the most robust probabilistic forecasts and their uncertainties. This is the approach we follow in the analysis below. They showed that the probability of another Carrington-like event within the next 10 years is 10.3% with confidence interval [0.9, 18.7].

Riley and Love (2017) also considered how the forecast might vary within a solar cycle. In some sense, this admits that the assumption of time stationarity breaks down. However, there are not a

sufficient number of data points in any of the datasets to look at limited portions within a single cycle (during which one can assume stationarity). To address this, they extracted data ±2.5 years around each solar minimum and each solar maximum and combined them into two new datasets, one representing solar minimum-like and the other solar maximum-like conditions. This was possible because of the assumption that the events are independent of one another. They found that the forecasts differed drastically between quasiminimum and maximum conditions. In particular, they estimated that the likelihood of another event as large, or larger than 850 nT, is 1.4% during solar minimum conditions and 28% for solar maximum conditions.

In this chapter, we review the basic techniques for estimating the likelihood of an extreme space weather event, focusing both on the estimate itself as well as the uncertainties associated with it, which, we believe, is equally important. We add to previous studies by estimating the probability of an extreme event based on the auroral electrojet *AE* index, and extend our previous *work investigating the Dst* index to also include the *Dst* proxy parameters *Dxt* and *Dcx*, which extend significantly further back in time. Additionally, we refine our previous analysis of SPEs by applying a more rigorous approach.

2 METHODOLOGIES

In this section, we summarize the primary datasets used in our analysis as well as the statistical modeling approach used to make the probabilistic forecasts of extreme events.

2.1 DATASETS

Predicting an extreme space weather event first requires that we define what one is. This in turn means that we must choose parameters that describe events that are distributed in some measure of severity. As a concrete example, we could look at the speeds of CMEs, which vary from a few hundred kilometers per second to almost 4000 km s^{-1}. These speeds are not distributed normally, but have a power-law or log-normal distribution (Riley and Love, 2017). We can arbitrarily define "extreme" events as those that have speeds, $v > 3000$ km s^{-1}.

For continuous datasets, on the other hand, we must define "events" within the data stream. The *Dst* index, for example, is a continuous time series. Magnetic storms are identified as intervals of depressed *Dst* over an extended period of time. We can capture the storm by identifying the minimum value within the event.

Previously, Riley (2012) studied a broad range of datasets, including solar flare data, CME speeds, the *Dst* index, and SPEs as determined from nitrate data, discussing the relative merits and limitations of each. In this chapter, we focus on the *Dst* index and two proxy measures for *Dst*, *Dxt*, and *Dcx*; the auroral ejectrojet index, *AE*; and a measure of SPEs (specifically, all >30 MeV proton events with fluences exceeding 10^9 pr cm^{-2}). These add to the data previously analyzed but, also for data already analyzed, extend the methodologies applied. This is obviously not an exhaustive list. We could, for example, consider CME mass and energy, IMF Bz intervals, and the equatorward edge of the diffuse aurora, to name a few. Analysis of these, however, should probably be driven by need. Kp is an obvious index that would benefit from such an analysis; however, because it only takes on a limited number of values, it is not clear how to apply these techniques to such a parameter.

To complement the *Dst* index, we also consider the *Dxt* and *Dcx* indices (Karinen and Mursula, 2006). These indices are attempts to correct defects in the *Dst* index, and, additionally extend the duration of the index back in time to 1932. *Dxt* is considered to be a reconstructed Dst index. It correlates highly with *Dst* (0.987 for hourly values) and, additionally, corrects errors present in the original *Dst* index (Karinen and Mursula, 2005). *Dcx* has been proposed as a corrected version of the *Dst* index to account for seasonally varying quiet-time levels that can raise $|Dst|$ by as much as 44 nT for an individual storm (Karinen and Mursula, 2006). The corrections appear robust in that they improve the index's correlation with sunspots and other geomagnetic indices. For our purposes, the main advantage lies in: (1) the extended duration of the dataset (80.1 years), allowing us to test the robustness of our forecasts over this extended period; and (2) the impact of changes in the peak values for severe storms. For example, the two notable storms of March 23, 1940 ($Dcx/Dxt_{min} = -355/-360$ nT) and September 18, 1941 ($Dcx/Dxt_{min} = -404/-417$ nT) are contained within the datasets, but not within the *Dst* dataset. Of course, this comes with an important caveat that a storm value of -850 nT in *Dst* might not be the same as -850 nT in *Dxt* or *Dcx*. Additionally, it is worth emphasizing that these data may contain artifacts not present in the more comprehensive and well-studied *Dst* dataset. For example, during the September 18, 1941 storm, the light trace at the Honolulu observatory dropped off the edge of the photographic paper for 10 hours, suggesting that the measurement "saturated." However, a data gap is not noted in the *Dcx* data, nor is it readily apparent from an inspection of the time series. The errors introduced by such "saturation" of the dataset remain to be analyzed, estimated, and, if possible, mitigated. Nevertheless, with these caveats in mind, we believe the analysis of these complementary datasets is justified.

Additionally, we also investigate the properties of the *AE* index. The *AE* index was designed to provide a global estimate of auroral zone magnetic activity, driven by ionospheric currents both below and within the auroral oval (Davis and Sugiura, 1966). In essence, it is the amount by which the horizontal magnetic field around the auroral oval deviates from quiet-time values. In practice, calculating the *AE* index is a relatively complex process involving measurements of 10–13 observatories situated along the auroral zone, the normalization of these data to base values that are computed from quiet days, and the computation of two intermediary indices, known as AU and AL. Thus, while recognizing that the index is not a simple fiduciary for some simple physical variable, it does provide a measure for the overall activity of the electrojets, which, in turn, demonstrates its usefulness as a measure of substorm activity. The analysis of *AE* complements that of *Dst*, which relies on measurement stations located at low latitudes and, hence, minimizing the effects of the auroral zone.

There are several caveats that should be kept in mind when using the *AE* index (Mandea and Korte, 2010). First, diurnal variations might be present due to the limited latitudinal extent and uneven longitudinal distribution of the *AE* stations used to derive the index. Second, when the number of available stations falls below 12, there is the possibility that ejectrojet enhancement will be missed. Third, although the index is primarily driven by ionospheric electrojet current, magnetospheric currents (e.g., the equatorial ring current) may sometimes contribute to AE, complicating its interpretation.

Finally, we apply our techniques to a dataset derived from nitrate records in polar ice cores, which, at least to some, are believed to capture SPEs. It stretches back more than 400 years, and thus represents the only dataset within which the actual Carrington event is embedded. In principle, they are a measure of the flux of a population of highly energized particles, accelerated either by the flare or CME-driven shock associated with an extreme event. However, they are not without controversy or caveats. First, while the nitrate spikes were, until recently, believed by space physicists to be a record of large,

historical space weather events (McCracken et al., 2001), ice-core chemists are skeptical (Wolff et al., 2008). They posit that no viable mechanism exists by which SPEs could be imprinted within the ice, suggesting instead that high concentrations of sea salt provide a simpler and more consistent explanation for the deposition of aerosol nitrates. Additionally, recent work by space physicists also casts doubt on their solar origin (Duderstadt et al., 2016). Second, there are only 70 events spanning the 450 years for which we have data. The largest event in the dataset, with a fluence of 18.8×10^9 cm^{-1}, occurred in 1859. That is, the largest event in the last 400 years was the Carrington event. More importantly, however, with such a limited number of events, the statistics of the fit and the resulting probability estimates will be more prone to error. In spite of these limitations, and under the assumption that they are capturing extreme solar phenomena, we can estimate their likelihood of occurrence.

2.2 STATISTICAL MODELING

Here, we review a method for estimating the likelihood of another Carrington-like event by assuming that the events are distributed in severity in a way that can be described by a continuous curve (e.g., exponential, log-normal, or power-law). However, several other approaches could—at least superficially—be used, including event trees, similarity judgments, and time to event. However, for the reasons outlined by Riley (2012), none of these approaches are amenable for studying extreme space weather phenomena. Thus, we focus on an extrapolation technique that assumes we can continue the curve of well-observed events out into the region of frequency-severity space for which there are few or even no observed events.

Riley (2012) showed that a range of space weather phenomena can—at least qualitatively—be described by a quasipower-law distribution. To generalize this, and test whether another distribution is more appropriate, we consider the following three types of distributions: power-law (PL), log-normal (LN), and exponential (E). To this, we can add so-called "cut-off" distributions where the data dramatically drop off at some point.

First, a set of events, x, obeys a power-law (or Pareto) distribution if the probability of occurrence, $p(x)$, can be written:

$$p(x) = C_1 x^{-\alpha} \tag{1}$$

where the exponent, α, is a fixed parameter, and C_1 is estimated from the location at which the curve intercepts the y-axis. Similarly, a set of events, x, is said to follow a log-normal distribution if the probability of occurrence, $p(x)$, obeys:

$$p(x) = \frac{C_2}{x} e^{-(\ln x - \mu)^2 / 2\sigma^2} \tag{2}$$

where μ and σ are parameters that must be fit based on the observations, and C_2 is another constant. Finally, a set of events, x, is said to follow an exponential distribution if:

$$p(x) = C_3 e^{-\lambda x} \tag{3}$$

where λ is a free parameter whose value is estimated based on a best fit to the measurements, and C_3 is another constant. This triplet of distributions, we believe, reasonably encompasses the relevant phase space with power-laws falling off least quickly, exponentials falling off most rapidly, and log normals generally lying between the two.

2.2.1 Estimating the best-fit parameters to a model

Riley (2012) described a technique for estimating the likelihood of a space weather event for power-law distributions, based on earlier work by McMorrow (2009). Following this, we define the complementary cumulative distribution function (CCDF), $P(x)$, as the probability of an event of magnitude equal to or greater than some critical value x_{crit}:

$$P(x \geq x_{crit}) = \int_{x_{crit}}^{\infty} p(x')dx' \tag{4}$$

which, for a finite dataset, simplifies to:

$$P(x \geq x_{crit}) = \frac{C}{\alpha - 1} x_{crit}^{-\alpha+1} \tag{5}$$

Here, the CCDF also obeys a power law with a lower exponent ($\alpha - 1$). CCDFs offer a number of benefits over the original power-law distributions: (1) they circumvent issues associated with noisy tails; (2) the slope can be computed using the maximum likelihood estimate (MLE):

$$\alpha - 1 = N \left[\sum_{i=1}^{N} \ln \frac{x_i}{x_{min}} \right]^{-1} \tag{6}$$

where x_i are the measured values of x, N is the number of events in the dataset, and x_{min} is some appropriate minimum value of x, below which the power-law relationship breaks down (Newman, 2005); and (3) the CCDF naturally generates the probability of occurrence of some event of a particular strength or greater, not the probability of an event of size x.

Using Eq. (5) we can estimate the number of events as large as or larger than x_{crit} during the period covered by the dataset, E:

$$E(x \geq x_{crit}) = NP(x \geq x_{crit}) \tag{7}$$

where N is the total number of events within the dataset.

Finally, again under the assumption that the events happen independently, we can employ the Poisson distribution to derive the probability of one or more events greater than x_{crit} occurring during some time Δt:

$$P(x \geq x_{crit}, t = \Delta t) = 1 - e^{-N \frac{\Delta t}{\tau} P(x \geq x_{crit})} \tag{8}$$

where τ is the total time span of the dataset. Eqs. (5), (6), (8) thus provide a robust technique for calculating the probability that an event of severity exceeding x_{crit} will occur some time within the next Δt years.

Similar expressions can be written for log-normal and exponential distributions, and Eq. (8) can be used to estimate probabilities based on these distributions.

2.2.2 Identifying the tail in the distribution

It is unlikely that natural phenomena display tail-like behavior throughout their entire distribution. At the lowest frequencies, saturation effects likely dominate. Similarly, at the highest frequencies, a cut-off must be anticipated at some (even remote) point, based on, say, physical constraints (e.g., maximum possible available energy). Thus, we need to identify a lower limit in severity, above which we can reasonably argue that a tail-like distribution exists.

Riley et al. (2012) and Love et al. (2015) identified the minimum values in severity (x_{min}) manually, and, arguably, somewhat subjectively. In particular, Riley et al. (2012) chose $x_{min} = 120$ nT while Love et al. (2015) used $x_{min} = 63$ nT as a lower bound for the $|Dst|$ index. Here, we use an approach for optimizing the value of x_{min} based on minimizing the Kolmogorov-Smirnov (KS) goodness-of-fit statistic between the model and the data (Kolmogorov, 1933; Smirnov, 1948; Press, 2007). Essentially, the KS statistic aims to balance the inclusion of tail data to improve sample statistics while at the same time omitting low-severity data that may not reflect the true nature of the tail.

2.2.3 Nonparametric bootstrapping

Following Efron and Tibshirani (1994), we apply a technique known as nonparametric bootstrapping to estimate the confidence intervals of our predictions. The observed data is randomly sampled and new pseudo-datasets are constructed from the events drawn (and replaced). Intuitively, this process can be understood as follows: You place all the data into a bag and randomly pick one out. You record the value, replace the datapoint back into the bag and choose another. To each of these, one of the three distribution profiles (PL, LN, or E) is fit. The bootstrap approach is straightforward to apply and generally provides reasonable estimates of standard errors and confidence intervals when the sample size is large.

Once a sufficiently large number of pseudo-datasets (i.e., bootstraps) have been computed and fit, the variability within these profiles can be used to define, say, 95% (i.e., between 2.5% and 97.5%) confidence intervals.

2.2.4 Model comparison

Using the techniques outlined thus far, we can use the computed bootstrapped fits to test whether a PL distribution is plausible by computing a p-value. We define the null hypothesis (H_0) to be that the power law adequately describes the data, and the alternative hypothesis (H_1), that some other distribution is better. Thus, if $p > 0.1$, say, then the difference between the data and the model can be attributed to statistical fluctuations and we cannot reject H_0. On the other hand, if p is small, say, <0.1, then the PL model is not a plausible fit to the data. It should be emphasized that merely because the p-value is large, this does not guarantee that the PL model is correct; for that we must apply a model comparison test.

Vuong's test is one such model comparison test that relies on a likelihood ratio test for selecting one model over another (Vuong, 1989). Specifically, it uses the Kullback-Leibler divergence, which is a measure of the difference between two probability distributions, say, A and B (Kullback and Leibler, 1951; Joyce, 2011). The criteria estimates the information gained or lost when model B is used to approximate model A. Alternatively it can be thought of as a metric that measures the distance between A and B.

In our case, we compute Vuong's test statistic, R_V, which compares two models under the hypothesis that both classes of distribution are equally far from the true distribution. If true, the log-likelihood ratio would have a mean value of zero. R_V moves toward $\pm\infty$ if one model is substantially better than the other. Additionally, one-sided and two-sided p-values can be computed to estimate the significance of the R_V statistic. The one-sided approach tests the null hypothesis (H_0) that both classes of distributions are equally far from the true distribution, against the alternative hypothesis (H_1) that model A is closer to the true distribution. The two-sided version tests the null hypothesis (H_0) that both classes of distributions are equally far from the true distribution, against the alternative (H_1), that one of the

distributions is closer. In both cases, we reject H_0 if $p < p_{crit}$, where in this case, we conservatively chose $p_{crit} = 0.05$.

For a complementary approach, including estimates of the probabilities for magnetic storms produced by different drivers (CIRs, sheath, MC, and CME), please see Chapter 4 in this volume.

3 RESULTS

In this section we apply the previously described techniques to *Dst*, *Dxt*, *Dcx*, *AE*, and SPE >30 MeV fluences.

Beginning with the $|Dst|$ index, in Fig. 1, we show a time series of severe ($|Dst| > 100$ nT) geomagnetic storms. We note several points. First, only one event approached 600 nT, and, moreover, only five events exceeded 400 nT. Second, the storms appear to cluster on the timescale of ~11 years, in effect mimicking but trailing the sunspot cycle. Third, bimodal peaks can be seen at and after solar maximum, matching CME rates (Riley et al., 2006). Fourth, there is a tendency for the strongest storms to become stronger from 1965 to 2007. In particular, while the largest five storms around 1970 were between 200 and 300 nT, the five most intense storms around 2005 were between 350 and 450 nT. On the other hand, the most recent decade shows a relative dearth of events, and a particular lack of intense storms. In summary then, there appear to be both periodic and secular variations in the time series.

Fig. 2 summarizes the probability estimates using the three possible distributions. The panel on the left (A) shows the CCDF, which, as discussed earlier, is the probability that an event as large or larger than some critical value will occur during a unit time interval. The open circles show all the events. The advantage of using the CCDF rather than the underlying $p(x)$ is self-evident: The data are not binned in x but rather summed so that the number of data points used to construct each open circle is the sum of all the data points to the right of itself (e.g., Riley et al., 2012). The points are well represented by a straight line at least up to ~280 nT. Beyond this, with the exception of the most severe storm, they appear to "fall off" this trajectory.

The colored curves show a selection of fits to the bootstraps. Specifically, for each of 1000 bootstrapped pseudo-datasets, a PL, LN, and E distribution was fit. Of these, 100 randomly chosen ones are displayed. The general conclusion, at least visually, is that: (1) the PL profiles capture the lower severity measurements but overestimate the likelihood of the most severe events; (2) the LN profiles underestimate the low-severity events but capture the trends at higher severity; and (3) the E profiles overestimate the low-severity frequency and underestimate the high-severity frequency of events.

Fig. 2B summarizes the likelihood of observing an event as severe or more severe as the most severe event observed (i.e., $Dst < -589$ nT) during the entire span of the data (~57 years). For a PL/LN/E distribution, the median probabilities are: 0.95, 0.63, and 0.13, respectively.

Using Eq. (8), we can estimate the probability of such an event occurring over the next decade to be 20.3/3.0/0.02% for a power-law, log-normal, or exponential distribution (see also Table 1). Moreover, we can use the bootstrap results to estimate confidence intervals in these predictions. For the power-law distribution, for example, our best estimate is 20.3% for 95%CI [12.5,30.2]. Table 1 also shows the forecast when only data from 1964 through the present is included in the analysis. In this case, estimates drops by almost a factor of two. Additionally, and not shown, if we require that an event exceeds a threshold of −1700 nT, a value closer to that suggested by Tsurutani et al. (2003) for the Carrington

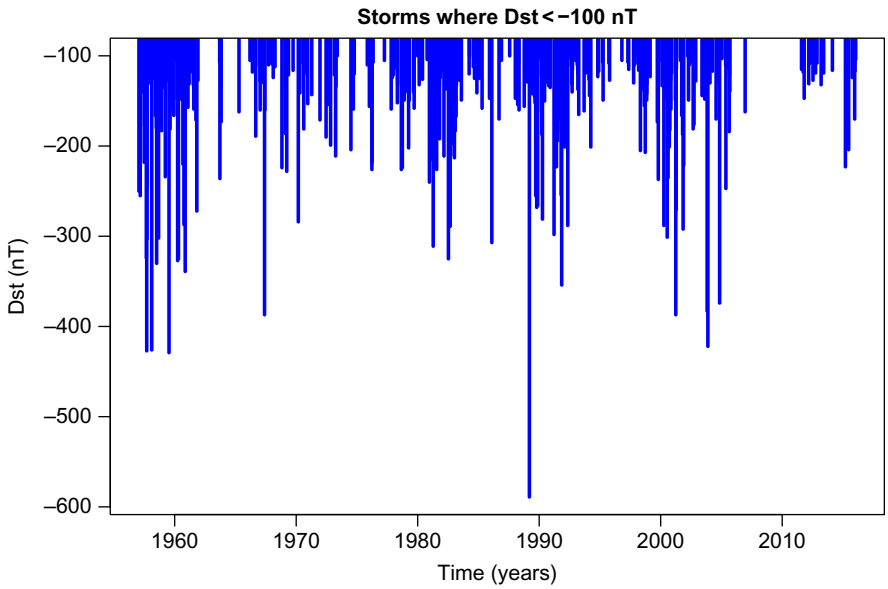

FIG. 1

Magnetic storms, defined by events where $|Dst|$ exceeds 100 T, are shown as a function of time. Individual storms are identified as contiguous intervals where $|Dst|$ exceeded 100 nT. The data were obtained from NASA's COHOWEB.

Based on Riley, P., Love, J.J., 2017. Extreme geomagnetic storms: probabilistic forecasts and their uncertainties. Space Weather 15, 53–64.

event, the probability of occurrence over the next decade decreases to 1.5%. This is quite comparable to the result obtained by Yermolaev et al. (2013), who estimated that the likelihood of a storm with $|Dst| >$ 1760 nT was not higher than one event every 500 years (or, a probability of ∼2% per decade).

Fig. 3 summarizes the main statistical parameters for the bootstrap fits. For each parameter, the cumulative mean and 25% and 75% quantiles are shown as a function of iteration (i.e., the number of bootstraps). Thus, as the number of bootstraps is increased, our estimate for the different parameters improves. The best-fit values are: $x_{min} \sim 123$ nT and $\alpha \sim 3.72$, and the number of points used to construct the tail statistics is ∼250.

Using the computed bootstrap fits, we can also test the hypothesis of whether the power-law distribution is plausible. The p-value for this is shown in Fig. 3D. Unlike the more usual approach for interpreting p-values, this one is set up such that a value <0.05 provides strong evidence against the power-law hypothesis. On the other hand, values above 0.05, or, more conservatively, above 0.1 would suggest that a power-law distribution is plausible. Thus, the value determined for the *Dst* index, ∼0.2, suggests that we cannot rule out the power-law as the underlying distribution. On the other hand, this result does not mean that this is the correct distribution. For that, we need to apply Vuong's test and make direct comparisons amongst viable models.

In Table 2 we compare Vuong's test statistic (R_V) for each of three comparisons: Power-law vs. log-normal; power-law vs. exponential; and log-normal vs. exponential. Additionally, we provide p-values for both the one-sided and two-sided tests. Larger positive values of R_V provide support

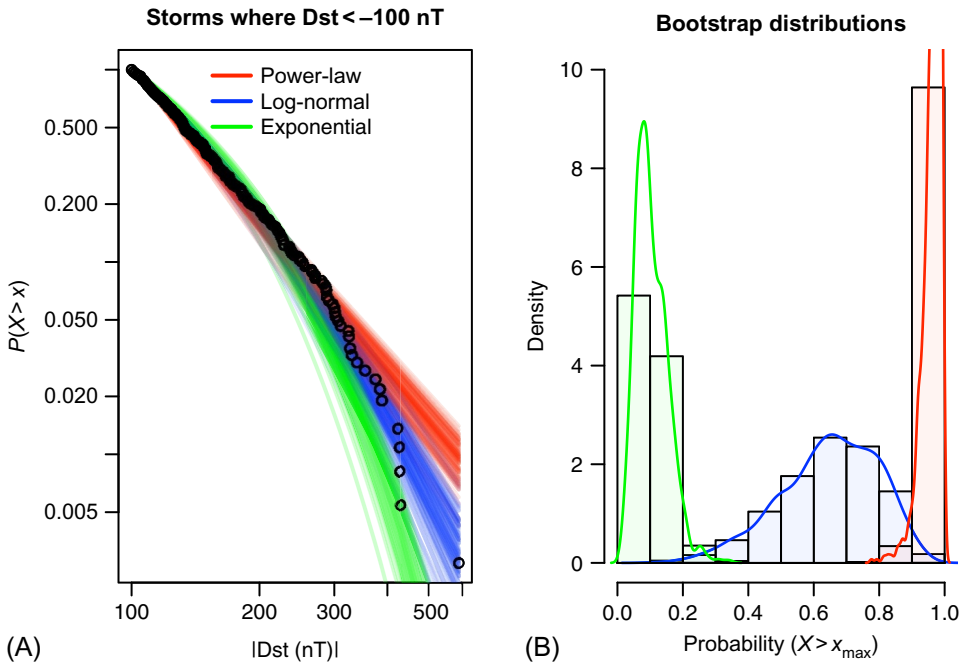

FIG. 2

(A) Complementary cumulative distribution function (CCDF) for the geomagnetic storms shown in Fig. 1. Bootstrap fits for the three distributions (power-law, log-normal, and exponential) are superimposed. (B) Histogram and density plots showing the probability of an event as large as or larger than the largest event in the dataset (−589 nT) over the duration of the dataset (~49 years). The density curve colors follow the convention given in the legend within panel (A).

Table 1 Best Estimates and Confidence Intervals for *Dst*

Distribution	Median (%)	2.5% (%)	97.5% (%)
Power-law (1964–2016)	10.3	0.9	18.7
Power-law (1957–2016)	20.3	12.5	30.2
Log-normal	3.0	0.6	9.0
Exponential	0.02	0.004	0.08

for the first model over the second model. Thus, we can infer from these results that the log-normal distribution is slightly favored against the power-law, and much more so against the exponential distribution. Similarly, the power-law is favored over the exponential. However, for these results to be statistically significant, we require p-values < 0.05. In almost all cases, the calculated p-values exceed this, and, in most cases, substantially so. Only the two-sided log-normal vs. exponential p-value is less than this threshold (0.02), which would allow us to firmly discount the exponential distribution.

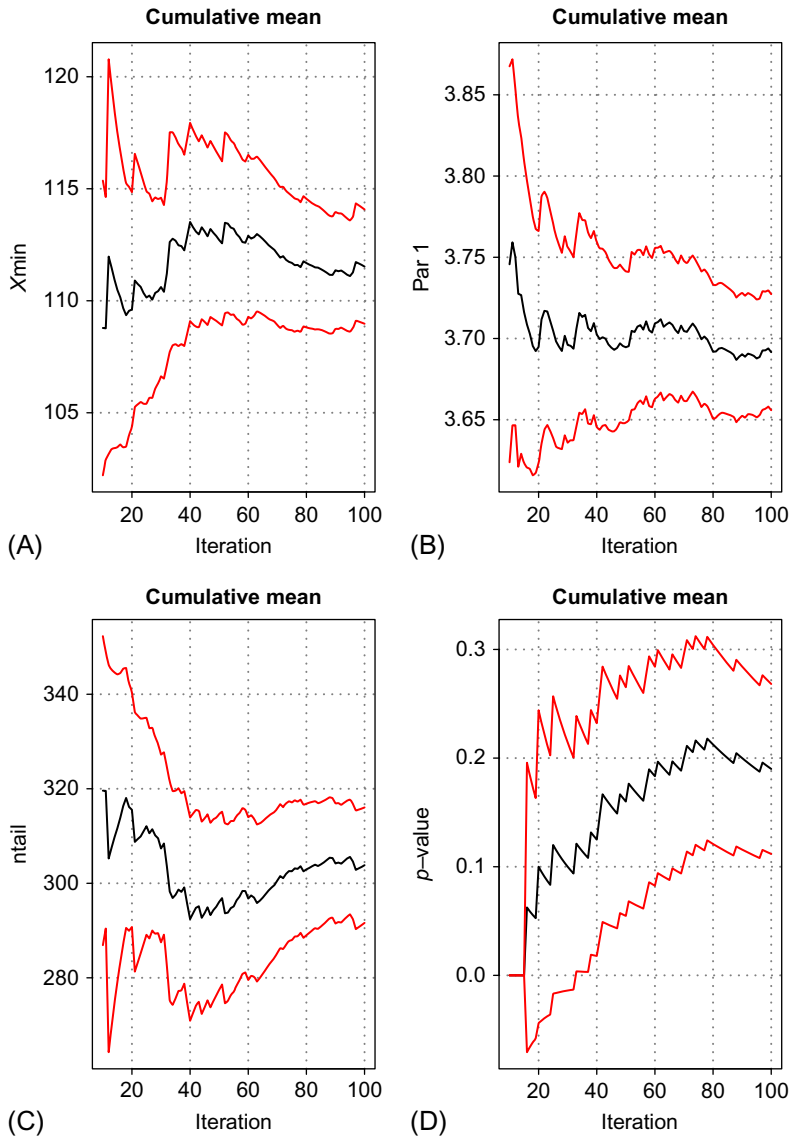

FIG. 3

Summary of statistical parameters for power-law bootstrap fit to the *Dst* dataset as a function iteration (i.e., the number of bootstraps). (A)–(D) The cumulative means of (A) x_{min}, (B) α (Par1), (C) n_{tail}, and (D) the p-value.

Table 2 Vuong's Test Statistics for the *Dst* Index, Where P-L, L-N, and Exp. Refer to Power-Law, Log-Normal, and Exponential Distributions

Statistic	P-L vs. L-N	P-L vs. Exp.	L-N vs. Exp.
R_V	−0.488	1.08	2.29
One-sided	0.313	0.86	0.989
Two-sided	0.626	0.279	0.0219

Table 3 Best Estimates and Confidence Intervals for the *Dst* Index for Each Solar Cycle From 1957 Through Early 2016

Cycle	Interval	Power-Law	Log-Normal
19	57–64	65.02 [27.54, 90.67]	16.45 [1.05, 46.45]
20	64–76	0.15 [1×10^{-4}, 4.76]	2.5×10^{-12} [0, 1.17]
21	76–86	14.34 [3.36, 37.50]	0.14 [1×10^{-4}, 3.74]
22	86–96	0.044 [4.2×10^{-11}, 4.33]	1.8×10^{-7} [0, 0.1]
23	96–08	12.82 [3.89, 30.27]	3.21 [0, 14.0]
24	08–16	0.049 [2.4×10^{-9}, 4.43]	2×10^{-7} [0, 0.1]

3.1 ASSESSING THE VALIDITY OF THE TIME STATIONARITY ASSUMPTION

The approach adopted here has relied on the assumption that the data are time stationary. As we have discussed, however, there are both cyclic and secular variations in space weather phenomena. To better understand the impact of this variability, we have repeated our analysis for each of five epochs: Solar cycles 19, 20, 21, 22, and 23/24, which cover the time period from 1957 through 2012. Breaking the data into these five intervals necessarily increases the uncertainties associated with any predictions we make. Nevertheless, it may, in principle, provide some information about the intrinsic variability from one cycle to another. Because the number of events following the end of cycle 23 was so small, we grouped cycle 24 events with cycle 23.

Table 3 summarizes the probabilities estimated using both the power-law and log-normal distributions, using the same analysis as described above. We note the following points. First, there is considerable variability from one cycle to the next, suggesting that either (1) time stationarity is not a reasonable approximation; or, and more likely, (2) the limited sample size for a single decade is not large enough to compute a meaningful estimate of the probability. Due to the significant scatter, we cannot discern any obvious secular trend in these probabilities. On the other hand, with the exception of the 57–64 interval, an estimate of ∼4% per decade is contained within all confidence bounds. Additionally, when estimates are made for the intervals 1957–2016 and 1964–2016 (Table 1), the former produces significantly higher forecasts. Thus, we suggest that the interval from 1957 to 1964 was indeed associated with a significantly larger probability for an extreme event.

3.2 ANALYSIS OF *Dxt* AND *Dcx*

Fig. 4 summarizes the CCDFs for *Dxt* and *Dcx*. To a large degree, the profiles are quite similar to those produced using the original *Dst* record. When computing the 10-year forecasts for an event as large or larger than −850 nT using *Dxt* and *Dcx*, we find values of 17.7% and 15.9%, respectively, with confidence bounds that are comparable to those computed for the *Dst* index.

3.3 EXTREME SPACE WEATHER EVENTS IN THE IONOSPHERE: THE *AE* INDEX

Fig. 5 shows all *AE* "events" for which the index exceeded 1000 nT. We used the same procedure to identify "events" as with the *Dst* index. In particular, we identified all contiguous intervals for which the *AE* index exceeded 200 nT, then for each of these located the peak value and timing of that peak. We then discarded those events that did not exceed 1000 nT. Over the ∼58-year time span (from 1957 to 2012), there were 1165 events matching this criteria. Solar cycle modulation and any possible secular trends are much less pronounced in this dataset, although an argument can be made that the last eight years have been unusually quiet. The most severe *AE* "event" occurred on March 23, 1991, reaching a value of 3195 nT.

Fig. 6A shows the CCDF for all *AE* "events" that were larger than 1000. Bootstrap solutions for each of the three distributions are shown again in red (PL), blue (LN), and green (SE). The PL bootstraps are, at least subjectively, consistent with the observed tail: Arguably only two of the most severe events are not enveloped within the bootstrap spread. Similarly, at least visually, the LN distributions capture the fall off reasonably well. Only the E distributions fail to capture the profile correctly, dropping off too rapidly to be able to account for the largest events. It could be argued that x_{min} was not set correctly for these data. However, as we have discussed above, by relying on the KS statistic we have removed the subjectivity of fitting to less (the more severe) data in order to include more data and improve the statistics of the fit. Only with some extrinsic knowledge, such as there being an established break in the data, say, produced by different physical mechanisms, would we be justified in manually setting x_{min}. In summary, the data appear to trace the border between the PL and LN distributions.

Fig. 6B compares the probability distributions for predicting an event as large as or larger than, the largest event within the dataset (3195 nT). Focusing on the PL and LN distributions, we note the significant difference in estimates: A median probability of 0.67 or 0.34 depending on whether one chooses a PL or LN distribution, respectively. Table 4 summarizes these estimates and also provides 95% confidence intervals on these predictions.

The best-fit model parameters based on the bootstraps gave: $x_{min} \sim 1165$ nT and $\alpha \sim 7.16$. With a p-value of ∼0.45, we cannot reject the hypothesis that the data are well described by a power law. Additionally (not shown), we remark on a positive correlation between x_{min} and α: As x_{min} increases, so does α. This suggests that the slope of the curve is becoming increasingly steeper at higher values of AE, perhaps indicating that the underlying distribution is not best approximated by a straight line (i.e., constant α) but by a log-normal distribution, at least for the most severe events.

We also estimated Vuong's statistics associated with these fits. Again, the log-normal distribution was slightly favored over the power-law, and more strongly over the exponential. Similarly, the power-law was favored over the exponential. None, however, were statistically significant. Thus, based on statistical arguments alone, we cannot reject any.

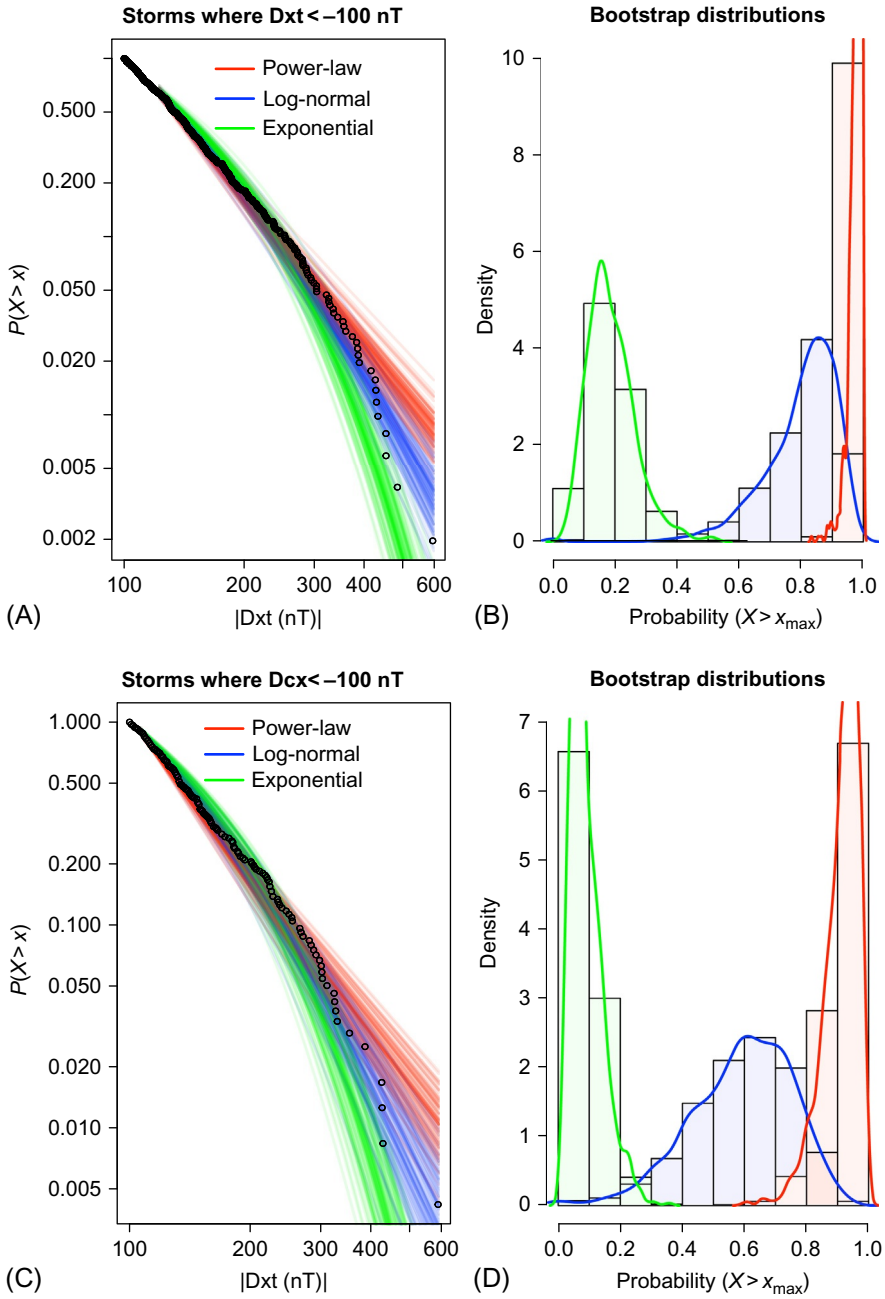

FIG. 4

As Fig. 2 but for *Dxt* (A and B) and *Dcx* (C and D).

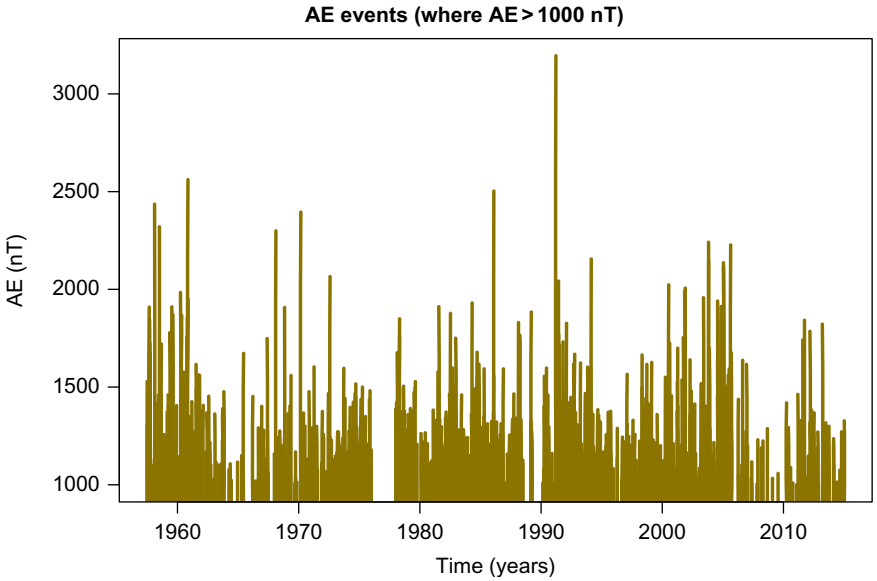

FIG. 5

Large auroral "substorms," defined by events where the *AE* index exceeds 1000 nT, are shown as a function of time. Individual storms are identified as contiguous intervals where *AE* exceeded 1000 nT.

Data from NASA's COHOWEB.

3.4 EXTREME SPACE WEATHER EVENTS IN THE HELIOSPHERE: ENERGETIC PROTONS

Fig. 7 shows all >30 MeV proton events with fluences exceeding 10^9 pr cm^{-2} as a function of time between 1562 and 1944. Although it is not possible to show rigorously because of the limited sample size, there is no obvious trend in the distribution of event sizes or temporal clustering to suggest that the time series is obviously nontime stationary. Of note is that the Carrington event is substantially larger than the other events in the dataset, with the second largest event producing a fluence of only 59% of the value of the 1859 event.

Fig. 8 shows the CCDF as a function of size and the probability density functions assuming power-law, log-normal, or exponential distributions. Specifically, the probability of observing an event as large as or larger than, the largest event observed in the dataset over a period as long as the interval over which the data were collected. Because the largest event contained within the nitrate record is presumed to be the 1859 event, this provides a more direct (albeit less reliable) estimate of the Carrington event. Thus, over the 382-year time span, the median probability of observing the Carrington event was 0.67/0.015/0.033 for the PL/LN/E distributions, respectively (Table 5).

The bootstrap fits, however, underscore the degree of uncertainty with these estimates. Based on KS statistics, only ~35 points were retained for the analysis. It is worth noting that the determination of x_{min} can be defended visually here. The shape of the curve appears to take three distinct slopes below $\sim 4 \times 10^{-9}$ cm^{-3}. Visually, at least, the log-normal and exponential distribution, which are overlaid

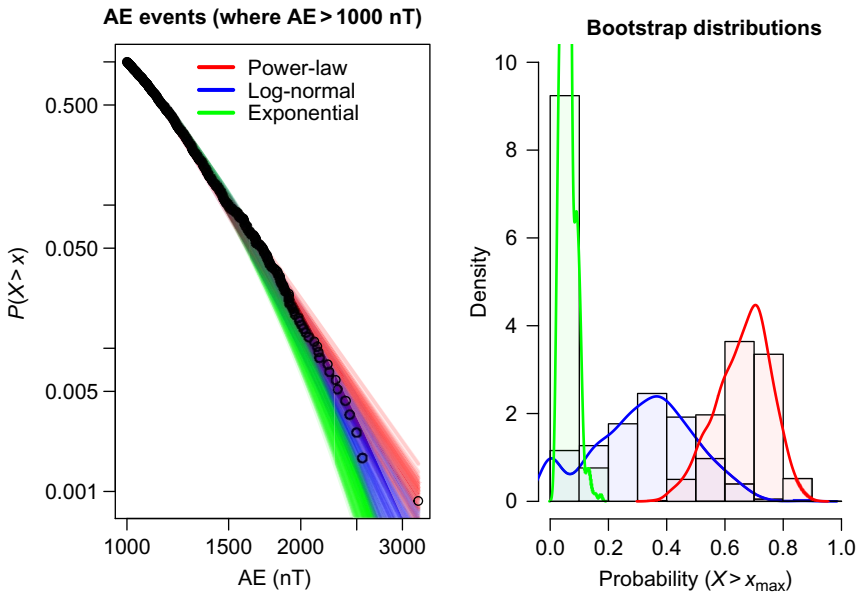

FIG. 6

As Fig. 1 for the auroral "substorms" shown in Fig. 5. Bootstrap fits for the three distributions (power-law, log-normal, and exponential) are superimposed. (*Right*) Histogram and density plots showing the probability of an event as large as or larger than, the largest event in the dataset (3295 nT) over the duration of the dataset (~51 years).

Table 4 Best Estimates and Confidence Intervals for AE

Distribution	Median (%)	2.5% (%)	97.5% (%)
Power-law	17.66	10.28	25.89
Log-normal	6.88	0.00	16.08
Exponential	1.08	0.40	2.08

upon one another, appear to capture the profile best, with the exception of the most severe (Carrington) event, which falls directly in the middle of the PL bootstrap fits. Application of Vuong's test this time slightly favors the log-normal over the exponential, and to a smaller extent the log-normal over the power-law. In turn, the power-law distribution is ever-so-slightly favored over the exponential distribution. Again, however, the comparisons are not statistically significant, using either the one-sided or two-sided p-values.

4 DISCUSSION

Although the analysis presented here has incorporated more statistical methodologies than our previous studies (Riley et al., 2012; Love et al., 2015), somewhat paradoxically, our analysis produced results that are even less constrained. Because we cannot be confident of the underlying distribution, and,

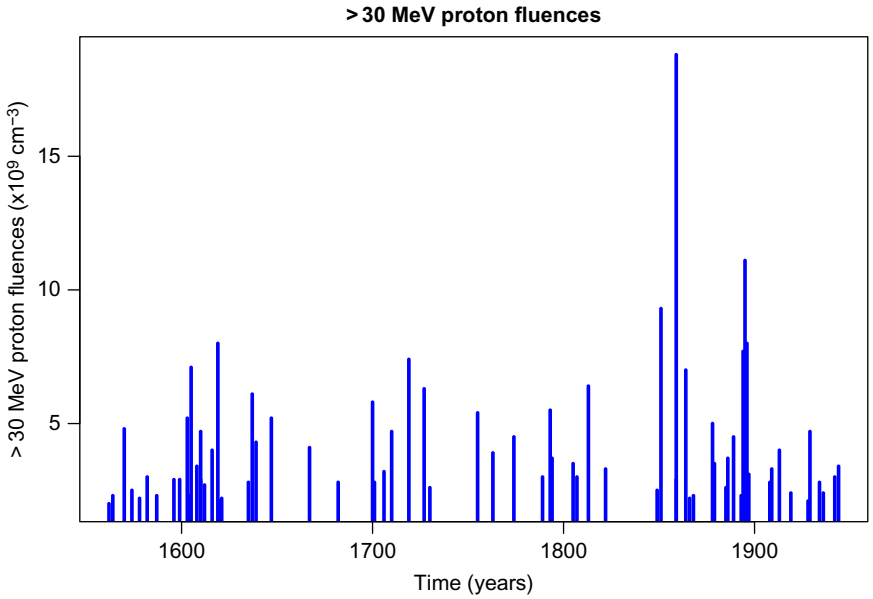

FIG. 7

Large SPEs, as defined by their inferred >30 MeV proton fluence exceeding 10^9 cm^{-3}, are shown as a function of time.

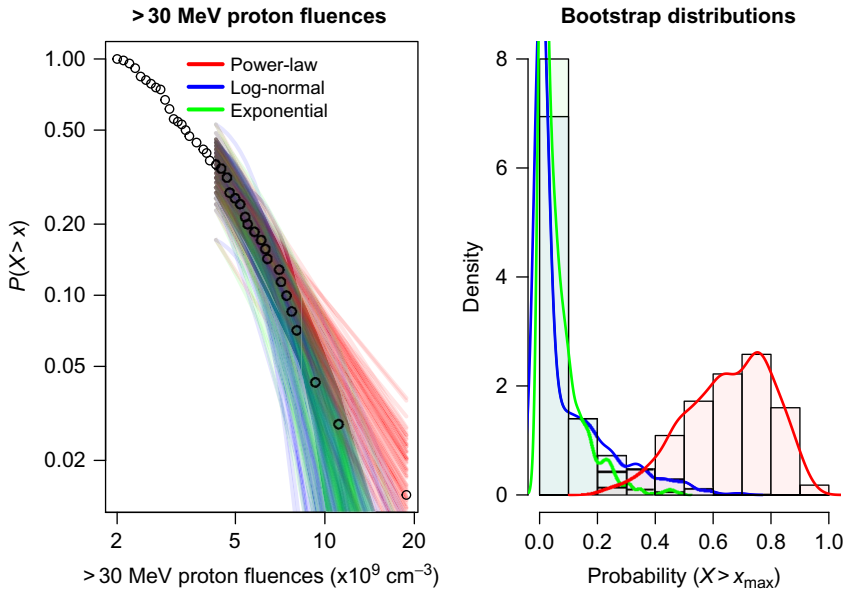

FIG. 8

As Fig. 1 but for the energetic SPEs shown in Fig. 7. Bootstrap fits for the three distributions (power-law, log-normal, and exponential) are superimposed. (*Right*) Histogram and density plots showing the probability of an event as large as or larger than the largest event in the dataset (18.8×10^9 cm^{-3}) over the duration of the dataset (~382 years).

Table 5 Best Estimates and Confidence Intervals for SPEs

Distribution	Median (%)	2.5% (%)	97.5% (%)
Power-law	2.85	1.06	5.67
Log-normal	0.04	1.5×10^{-9}	1.61
Exponential	0.086	0.0022	0.78

indeed, it appears that the data lie firmly between the power-law and log-normal distributions, our basic estimate could be as low as 3.0% or as high as 20.3%. While averaging these two values (11.65%) would yield a value that was remarkably similar to our two previous estimates (11% and 12%) (Riley et al., 2012; Love et al., 2015), in reality, it cannot be both. Statistical inference, however, can only be used up to a point. Beyond this we must use our best judgment, incorporating all relevant information to arrive at an informed estimate with credible confidence intervals. To do this, however, we must first address several important assumptions.

In our analysis, we assumed that the data are time stationary. At both ends of the temporal spectrum, we have shown that this approximation breaks down. The solar cycle modulates many solar parameters on the time scale of a decade or so (e.g., Riley et al., 2000). In particular, the largest 2% of geomagnetic storms (the so-called "super storms"), typically occur just after solar activity maxima (Bell et al., 1997). Thus, the dataset used to make forecasts should be at least this long, and any predictions made must necessarily be solar-cycle averaged estimates. On the other hand, we showed that forecasts based on data from solar minimum intervals might be as low as 1.4%/decade, but as high as 28% during solar maximum. It is interesting to note that this statistical result does not support the anecdotal view held by many in the scientific community that the most extreme storms tend to occur at or near solar minimum (based on the two extreme events of 1859 and 2012). It might, however, be possible to incorporate this knowledge into the prediction. For example, we might anticipate that the probability of an extreme event might peak a year or two after solar maximum, in conjunction with the peak in the rate of CMEs.

Similarly, over longer timescales, there is ample evidence for nonstationarity (e.g., Riley et al., 2012). In particular, Lockwood et al. (2009) inferred that there was a 10% likelihood that the Sun will fall into another grand minimum configuration over the next 40 years or so. If further substantiated, this information could be convolved with the current predictions for an extreme event on the premise that such phenomena would be significantly less likely to occur during Maunder-minimum-like conditions (Riley et al., 2015). Although speculative due to the absence of significant sunspots and, hence, active regions, there would be far fewer CMEs that could drive geoeffective space weather. Indeed, if the nitrate records in fact provide estimates of large SPEs, the Maunder minimum period lasting from ~1645 to ~1710 was the most quiescent period during the entire ~400-year interval. As a rough approximation, we can compute the frequency of space weather events during the Maunder minimum interval and compare that with a 65-year window centered on 1900, say, and use that ratio to adjust our forecasts based on the *Dst* index. There were seven events between 1645 and 1710 (1647, 1667, 1682, 1700, 1701, 1706, and 1710), and 12 events between 1867.5 and 1930.5 (1889, 1893, 1894, 1895, 1896, 1897, 1908, 1909, 1913, 1919, 1928, and 1929), suggesting that the current forecasts might need to be lowered by 58% if we are entering a grand solar minimum.

Although our analysis of decadal subsets of the *Dst* index did not yield any systematic trend in forecasts for Carrington events, constructing two long-window datasets (1957–2008 and 1964 to early

2016) suggests that the latter window is associated with a substantially lower forecast. Although the two intervals are roughly comparable (~51 years), the active latter half of the 1950s/early 1960s was replaced with the unusually quiet period surrounding and following the 2008/2009 solar minimum (also known as the "Eddy" minimum). This is further supported by the analysis of the *Dxt* dataset. When divided equally, estimates for an event that exceeds 850 nT are: 26.1%/decade for the period 1932–1973, and 12.3%/decade for 1974 to early 2016. If this captures the overall trend in solar activity into the future, we would anticipate that the future rate of occurrence of extreme events will be notably less than that estimated from the full (60 years) dataset. Obviously, if such conditions do ensue, probabilistic forecasts for extreme events may decrease; however, not necessarily to society's benefit. During periods of very low activity, for example, radiation from galactic cosmic rays (GCRs) will be higher, posing larger risks for passengers and airline crew as well as avionics. And, while SEP events themselves may decrease, the consequences from the ones that are produced may be more severe (e.g., Barnard et al., 2011).

The second major assumption addressed in this study is whether the data are best represented by a PL, LN, or E distribution. We inferred that the LN and PL distributions were consistent with the *Dst*, *Dxt*, *Dcx*, and *AE* datasets. For SPEs, either an L-N or E distribution is consistent with the data. The statistical results described and interpreted here, however, provide no guidance on the underlying causes for observing such distributions. In fact, where statistical summaries are ambiguous, we can reasonably resort to any available theories that might tend to favor one distribution over another. Several studies have alluded to the idea that substorms in particular can be described by self-organized criticality (SOC) (Angelopoulos et al., 1999; Klimas et al., 2000), which provides a natural explanation for the presence of power-law distributions for geomagnetic indices. On the other hand, we must temper this with the caveat that SOC has been invoked, and in fact was developed precisely because the measurements are observed to follow a power-law. By extension, we could posit that, in analogy with the Abelian Sandpile model (Bak et al., 1987), the magnetotail becomes progressively loaded until some specific threshold is reached, then reconnects (presumably in the form of multiple injections/substorms) and produces the observed $|Dst|$ maxima. However, this is undoubtedly a simplistic interpretation of a considerably more complex system. Love et al. (2015) has argued that perhaps the act of combining smaller storms, which do not apparently follow a power-law distribution, with larger storms that do may result in a distribution that is better approximated by a log-normal distribution. Indeed, this may be the case as they included storms down to values exceeding $|Dst| > 63$ nT. When KS statistics are used to identify $x_{min}(= 122$ nT), our results are consistent with a power-law distribution (but not inconsistent with a log-normal distribution).

A related but distinct assumption about the distribution is that it extends into a region of severity that we have observed rarely, if at all. Clearly, this assumption must fail at both extremes of the severity spectrum. In the low-severity portion of the spectrum, the curve usually flattens because smaller events are less easily identified or measured. At the high-severity portion of the spectrum, several factors may be important. First, "small-number" statistics may produce a curve profile that veers away from what would otherwise be a straight line. However, it is worth considering that even if the fluctuations at the extreme of the tail are random, we would expect a bias toward the undersampled region of this phase space, because the errors would not be expected to be distributed normally in log-log space. Second, in any finite-sized system, and particularly when the power-law distribution falls off more rapidly than higher frequency rates would suggest, the cut-off may be a real physical limitation. In practice, there must be a cut-off at some point. The key issue is whether that cut-off is near to, or far from, the critical

event under consideration. If the latter, we do not have to modify our analysis. However, if the former, we must account for the fact that events larger than the cut-off cannot contribute to our integrated estimate of events as large as, or larger than, some threshold. If we do not account for this, our estimates will be inflated.

For the *Dst* index, we can inquire what possible limits there might be. The absolute limit for $|Dst|$ is approximately 31,000 nT, which represents the complete cancelation of the Earth's magnetic field at the equator. Vasyliūnas (2011) has argued that the limiting value is considerably lower: ∼2500 nT. To arrive at this estimate, he set the plasma pressure equal to the magnetic pressure of the dipole field at the equator of each flux tube. This suggests a strong earthward gradient of the plasma pressure, which, through the relation $\mathbf{J}_{\perp} \sim (\mathbf{B} \times \nabla)\mathbf{B}/B^2$, implies a strong westward current through the magnetosphere. Vasyliunas used the Dessler-Parker-Sckopke relationship (Dessler and Parker, 1959; Sckopke, 1966) to arrive at the 2500-nT limit as a physical cut-off for *Dst*. Adopting this value would only marginally reduce the forecasts for a power-law distribution. In particular, using the full range of data from the Kyoto Observatory, the probability of an event as large or larger than 2500 nT is ∼1.3%. Thus, our estimate of 20.3% would only be reduced to ∼19.0%.

An interesting but as yet unexplored possibility is that if we could: (1) provide firm limits to the cut-off, and (2) argue for a log-normal distribution, this would allow us to fix the right-most portion of the curve, which, in turn, would allow us to better constrain the fit to the data and, hence, provide more accurate forecasts.

Our current forecast for an extreme event, where $Dst < -850$ nT, is 20.3%, assuming the data are distributed according to a power law, which is larger than two earlier estimates of ∼12% (Riley et al., 2012) and ∼11% (Love et al., 2015), although certainly within the overlapping confidence intervals. The disagreement between the current value and that in Riley et al. (2012) is due primarily to the addition of data from 1957 to 1963, which was a period of relatively high solar activity, and disproportionately added more high-severity storms to the data being analyzed. The disagreement with Love et al. (2015) is due to their use of (1) a lower, hand-picked value for x_{min} and (2) use of a log-normal distribution. In particular, the incorporation of the low-severity events strongly influenced the fit of the log-normal curve producing higher forecast estimates than would have been produced with a larger value of x_{min}. Thus, it is worth reemphasizing just how sensitively these results depend on the dataset under study as well as the techniques used to analyze them, and, as was pointed out by Love (2012), it is that it is not only the forecast estimates that must be communicated, but also the uncertainties and assumptions that accompany them.

Estimating the likelihood for future extreme space weather events can be of considerable value to decision makers. However, effectively communicating this information can be difficult. Probabilistic estimates with associated uncertainties can be phrased in any number of ways. Based on these results, for example, the likelihood of another extreme event on the scale of the Carrington or July 23 event over the next 12 months is only 2.3%. On the other hand, that same event has a 90% probability over the next 100 years. More importantly, while our study of extreme space weather events is important in its own right, it is perhaps the relative risk of a Carrington event as compared with, say, another earthquake on the scale of the 1906 San Francisco event or Hurricane Katrina, that is of more value to policy makers. Current 30-year probability estimates of an earthquake in California as large as or larger than, magnitude 8 are 4% (Field and Milner, 2008). Our estimate for an extreme space weather event on the scale of or larger than the "Carrington" event over the next decade is ∼20%. The 30-year estimate is 50%/decade (PL distribution) or 10% (LN distribution), which are 12 and 2.5 times larger than the

earthquake forecast. But even these comparisons can be misleading because it is the consequences of each disaster that society cares more about.

We should distinguish between natural disasters from natural hazards. The former refers to the actual event, such as a geomagnetic storm, flood, earthquake, or avalanche. It is the interaction between the disaster and society that makes a disaster a natural hazard (e.g., Blackwell, 2014). Practically speaking, the distinction is clear: estimating the damages caused by a disaster is more important than accurately describing the characteristics of the disaster itself. This is the information that both government agencies and insurance actuaries are primarily concerned with. Since there are often scaling laws between observables (e.g., magnitude earthquakes) and related physical parameters (e.g., earthquake rupture areas), we can infer the probability distributions of one from the other. Thus, we suggest that probabilistic forecasts can be significantly more informative when coupled not only with the magnitude of the related damages, but also with the likely costs associated with mitigating those damages. And, as noted earlier, these should be compared across multiple catastrophes allowing policy makers to allocate limited resources most effectively.

5 FUTURE STUDIES

A number of opportunities exist to build upon this work. First, constraining the uncertainties in the forecasts must be a core objective. Without tighter constraints, it is difficult for policy makers to make informed decisions about how to spend limited resources. But, to constrain the forecasts further is difficult. The spread in the bootstraps represents a baseline noise that cannot be removed with a limited number of events. Additionally, we cannot definitively establish which distribution best fits the data. If it were a log-normal distribution, the 10-year likelihood is a mere 4%. If it is a power-law distribution, it is 2.6 times higher (10.3%). Perhaps more sophisticated statistical techniques can be applied to delineate between the two? A further uncertainty concerns the assumption of time stationarity. Since the unusually quiet solar minimum of 2008/2009, the Sun has apparently entered a more quiescent state. Should this continue in the future, probabilistic forecasts based on space-era activity, which may have been anomalously high, would lead to erroneously large estimates for the 10-year forecasts. Instead, if the Sun were to maintain more solar-minimum-like activity, then the forecasts would be reduced by almost an order of magnitude (1.4%). Should we be entering into a more extreme solar wind state, such as a Maunder minimum (Riley et al., 2015), even this may be an overestimate. Thus, an interesting refinement to the techniques described here would be to develop estimates based on historical data, but which include a factor encapsulating what we think is a good estimate of solar activity over the next decade.

Our analysis has followed a purely statistical approach. However, physics-based modeling could be incorporated into the analysis. If, for example, firm upper limits for the size of extreme events could be found, that would modify the forecast estimates. As noted earlier, Vasyliūnas (2011) argued that the *Dst* index cannot exceed ~2500 nT. If so, this would reduce our estimate of 10.3% to 9%. If the physical limit were even lower, that would have a more substantial effect.

Another area that is ripe for analysis lies on the "consequences" side of the Sun-to-Earth chain. Can we use these statistical tools to estimate the probability of an electric field in the Earth's lithosphere exceeding some value? Or the current in the power grid? Or cosmic ionizing radiation within aircraft? In part, a lack of available data makes these questions difficult to address.

Statistical models can only answer carefully chosen questions in a limited way. We gain little in the way of a deeper understanding from such endeavors. Instead, they simply highlight a problem worthy of more careful investigation. This is where physics-based models can help. What are the solar wind parameters required to produced at *Dst* value of -1600 nT? What phenomena at the Sun and in the solar wind are required to generate such conditions? Knowledge of the underlying physical processes that produce extreme space weather events should ultimately result in the most accurate and useful forecasting tools. What made the July 23, 2012 extreme event so unique? Would we be able to recognize these or similar conditions in the future? And would we be able to extrapolate what we observe at the Sun to what would likely impact the Earth in the days to follow? This itself is a two-pronged question. First, given surface observations of an eruption, can we forecast the effects that will likely ensue at Earth? Second, and much more difficult, given a particular set of observations, can we reliably predict that there will be an eruption, and, if so, what the properties of that eruption will be at Earth?

6 CONCLUSIONS

In this chapter, we have described and applied a general technique for assessing both the likelihood and uncertainties of an extreme space weather event on the scale of or larger than, the Carrington event of 1859. In addition to the previously assumed power-law distribution, we also considered both log-normal and exponential distributions as alternatives for explaining the observed distribution in severity. Using the *Dst* index as the fiduciary measurement for defining a space weather event, we inferred that the probability of another event within the next decade exceeding 850 nT was 20.3% for 95% CI [12.5, 30.1] for a PL distribution but only 3.0% for 95% CI [0.6, 9.0] for an LN distribution. Increasing the threshold to 1700 nT reduced the estimate to 3.6% for 95% CI [1.7, 6.6].

Our studies thus far have not established whether a power-law or log-normal distribution best fit the data. We found that both are, within statistical uncertainties, consistent with the data. We also sought to establish tighter limits on our forecasts for the probability of another extreme event within the next decade; however, we found that depending on which datasets, intervals, and distributions were used to make the estimates, the results varied substantially.

In conclusion, under the assumptions that: (1) a PL distribution best represents the data; (2) the PL distribution is likely an upper limit to the behavior of the tail; (3) we are entering a period of lower solar activity; (4) a good definition of the *Dst* index for an extreme event is that $Dst < -850$ nT, we conclude that the best estimate for the probability of such an event over the next decade is approximately 10% for 95% CI [1, 20].

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NSF's FESD and NASA's LWS program, under which this work was performed. The Dst index was obtained from the Kyoto World Data Center (http://wdc.kugi.kyoto-u.ac.jp/dstdir/), while the *Dxt* and *Dcx* indices were provided by the Dcx server of the University of Oulu, Finland (http://dcx.oulu.fi).

REFERENCES

Angelopoulos, V., Mukai, T., Kokubun, S., 1999. Evidence for intermittency in Earth's plasma sheet and implications for self-organized criticality. Phys. Plasmas (1994–Present) 6 (11), 4161–4168.

Bak, P., Tang, C., Wiesenfeld, K., 1987. Self-organized criticality: an explanation of the 1/f noise. Phys. Rev. Lett. 59 (4), 381.

Baker, D.N., Li, X., Pulkkinen, A., Ngwira, C.M., Mays, M.L., Galvin, A.B., Simunac, K.D.C., 2013. A major solar eruptive event in July 2012: defining extreme space weather scenarios. Space Weather 11 (10), 585–591.

Barnard, L., Lockwood, M., Hapgood, M.A., Owens, M.J., Davis, C.J., Steinhilber, F., 2011. Predicting space climate change. Geophys. Res. Lett. 381, L16103. https://doi.org/10.1029/2011GL048489.

Bell, J.T., Gussenhoven, M.S., Mullen, E.G., 1997. Super storms. J. Geophys. Res. 102, 14189–14198. https://doi.org/10.1029/96JA03759.

Blackwell, C., 2014. Power law or lognormal? Distribution of normalized hurricane damages in the United States, 1900–2005. Nat. Hazards Rev. 16 (3) 04014024.

Davis, T.N., Sugiura, M., 1966. Auroral electrojet activity index AE and its universal time variations. J. Geophys. Res. 71 (3), 785–801.

Dessler, A.J., Parker, E.N., 1959. Hydromagnetic theory of geomagnetic storms. J. Geophys. Res. 64 (12), 2239–2252.

Duderstadt, K.A., Dibb, J.E., Schwadron, N.A., Spence, H.E., Solomon, S.C., Yudin, V.A., Jackman, C.H., Randall, C.E., 2016. Nitrate ions spikes in ice cores are not suitable proxies for solar proton events. J. Geophys. Res. 121 (6), 2994–3016.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC Press, Boca Raton, FL.

Field, E.H., Milner, K.R., 2008. Forecasting California's Earthquakes: What Can We Expect in the Next 30 Years? No. 2008–3027. U.S. Geological Survey.

Joyce, J.M., 2011. Kullback-Leibler divergence. International Encyclopedia of Statistical Science, Springer, pp. 720–722.

Karinen, A., Mursula, K., 2005. A new reconstruction of the Dst index for 1932–2002. Ann. Geophys. 23 (2), 475–485.

Karinen, A., Mursula, K., 2006. Correcting the *dst* index: consequences for absolute level and correlations. J. Geophys. Res. 111 (A8), 1–8.

Klimas, A.J., Valdivia, J.A., Vassiliadis, D., Baker, D.N., Hesse, M., Takalo, J., 2000. Self-organized criticality in the substorm phenomenon and its relation to localized reconnection in the magnetospheric plasma sheet. J. Geophys. Res. 105 (A8), 18765–18780.

Kolmogorov, A.N., 1933. Sulla determinazione empirica delle leggi di probabilita. Giorn. Ist. Ital. Attuari 4, 1–11.

Koons, H., 2001. Statistical analysis of extreme values in space science. J. Geophys. Res. 106, 10915–10921.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22 (1), 79–86.

Liu, Y.D., Luhmann, J.G., Kajdič, P., Kilpua, E.K.J., Lugaz, N., Nitta, N.V., Möstl, C., Lavraud, B., Bale, S.D., Farrugia, C.J., et al., 2014. Observations of an extreme storm in interplanetary space caused by successive coronal mass ejections. Nat. Commun. 5, 3481–3489.

Lockwood, M., Rouillard, A.P., Finch, I.D., 2009. The rise and fall of open solar flux during the current grand solar maximum. Astrophys. J. 700, 937–944. https://doi.org/10.1088/0004-637X/700/2/937.

Love, J.J., 2012. Credible occurrence probabilities for extreme geophysical events: earthquakes, volcanic eruptions, magnetic storms. Geophys. Res. Lett. 39. https://doi.org/10.1029/2012GL051431. L10301.

Love, J.J., Rigler, E.J., Pulkkinen, A., Riley, P., 2015. On the lognormality of historical magnetic storm intensity statistics: implications for extreme-event probabilities. Geophys. Res. Lett. 42 (16), 6544–6553.

Mandea, M., Korte, M., 2010. Geomagnetic Observations and Models. IAGA Special Sopron Book Series, Springer. https://books.google.com/books?id=DOMEIl7hlxsC. ISBN 9789048198580.

McCracken, K.G., Dreschhoff, G.A.M., Zeller, E.J., Smart, D.F., Shea, M.A., 2001. Solar cosmic ray events for the period 1561–1994: 1. Identification in polar ice, 1561–1950. J. Geophys. Res. 106, 21585–21598.

McMorrow, D., 2009. Rare events. JASON, The MITRE Corporation. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA510224&Location=U2&doc=GetTRDoc.pdf.

Newman, M., 2005. Power laws, pareto distributions and zipf's law. Contemp. Phys. 46, 323–351.

Press, W.H., 2007. Numerical Recipes: The Art of Scientific Computing, third ed. Cambridge University Press, Cambridge.

Riley, P., 2012. On the probability of occurrence of extreme space weather events. Space Weather 10 (null), S02012.

Riley, P., Love, J.J., 2017. Extreme geomagnetic storms: probabilistic forecasts and their uncertainties. Space Weather 15, 53–64.

Riley, P., Linker, J.A., Mikic, Z., Lionello, R., 2000. Solar cycle variations and the large-scale structure of the heliosphere: MHD simulations. In: IAU Joint Discussion, vol. 7. http://adsabs.harvard.edu/abs/2000IAUJD..7E.12R.

Riley, P., Schatzman, C., Cane, H.V., Richardson, I.G., Gopalswamy, N., 2006. On the rates of coronal mass ejections: remote solar and in situ observations. Astrophys. J. 647, 648–653. https://doi.org/10.1086/505383.

Riley, P., Lionello, R., Linker, J.A., Mikic, Z., Luhmann, J., Wijaya, J., 2012. Global MHD modeling of the solar corona and inner heliosphere for the whole heliosphere interval. Solar Phys. 274, 361–375. https://doi.org/10.1007/s11207-010-9698-x.

Riley, P., Lionello, R., Linker, J.A., Cliver, E., Balogh, A., Charbonneau, P., Crooker, N., DeRosa, M., Lockwood, M., Owens, M., et al., 2015. Inferring the structure of the solar corona and inner heliosphere during the maunder minimum using global thermodynamic magnetohydrodynamic simulations. Astrophys. J. 802 (2), 105.

Riley, P., Caplan, R.M., Giacalone, J., Lario, D., Liu, Y., 2016. Properties of the fast forward shock driven by the July 23, 2012 extreme coronal mass ejection. Astrophys. J. 819. https://doi.org/10.3847/0004-637X/819/1/57 57.

Russell, C.T., Mewaldt, R.A., Luhmann, J.G., Mason, G.M., von Rosenvinge, T.T., Cohen, C.M.S., Leske, R.A., Gomez-Herrero, R., Klassen, A., Galvin, A.B., Simunac, K.D.C., 2013. The very unusual interplanetary coronal mass ejection of 2012 July 23: a blast wave mediated by solar energetic particles. Astrophys. J. 770. https://doi.org/10.1088/0004-637X/770/1/38. 38.

Sachs, M., Yoder, M., Turcotte, D., Rundle, J., Malamud, B., 2012. Black swans, power laws, and dragon-kings: earthquakes, volcanic eruptions, landslides, wildfires, floods, and SOC models. Eur. Phys. J. 205, 167–182.

Sckopke, N., 1966. A general relation between the energy of trapped particles and the disturbance field near the Earth. J. Geophys. Res. 71 (13), 3125–3130.

Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. Ann. Math. Statist. 19 (2), 279–281.

Tsubouchi, K., Omura, Y., 2007. Long-term occurrence probabilities of intense geomagnetic storm events. Space Weather 51. https://doi.org/10.1029/2007SW000329. S12003.

Tsurutani, B.T., Gonzalez, W.D., Lakhina, G.S., Alex, S., 2003. The extreme magnetic storm of 1–2 September 1859. J. Geophys. Res. 108, 1268. https://doi.org/10.1029/2002JA009504.

Vasyliūnas, V.M., 2011. The largest imaginable magnetic storm. J. Atmos. Sol.-Ter. Phys. 73 (11), 1444–1446.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 307–333.

Wolff, E.W., Jones, A.E., Bauguitte, S.J.B., Salmon, R.A., 2008. The interpretation of spikes and trends in concentration of nitrate in polar ice cores, based on evidence from snow and atmospheric measurements. Atmos. Chem. Phys. 8 (18), 5627–5634. https://doi.org/10.5194/acp-8-5627-2008. http://www.atmos-chem-phys.net/8/5627/2008/.

Yermolaev, Y.I., Lodkina, I.G., Nikolaeva, N.S., Yermolaev, M.Y., 2013. Occurrence rate of extreme magnetic storms. J. Geophys. Res. 118 (8), 4760–4765.