AGU PUBLICATIONS

Space Weather

RESEARCH ARTICLE
10.1002/2016SW001470

Key Points:
- The probability of another extreme event is sensitive to the definition of "extreme"
- Likelihood of a Carrington event ($Dst < -850$) over the next decade is 10.3% 95% CI [0.9, 18.7]
- The uncertainties and assumptions used to derive forecasts are equally important

Supporting Information:
- Supporting Information S1

Correspondence to:
P. Riley,
pete@predsci.com

Citation:
Riley, P., and J. J. Love (2016), Extreme geomagnetic storms: Probabilistic forecasts and their uncertainties, *Space Weather*, *15*, doi:10.1002/2016SW001470.

Received 7 JUL 2016
Accepted 11 DEC 2016
Accepted article online 15 DEC 2016

Extreme geomagnetic storms: Probabilistic forecasts and their uncertainties

Pete Riley[1] and Jeffrey J. Love[2]

[1]Predictive Science, ,San Diego, California, USA, [2]U.S. Geological Survey, Geomagnetism Program, Denver, Colorado, USA

Abstract Extreme space weather events are low-frequency, high-risk phenomena. Estimating their rates of occurrence, as well as their associated uncertainties, is difficult. In this study, we derive statistical estimates and uncertainties for the occurrence rate of an extreme geomagnetic storm on the scale of the Carrington event (or worse) occurring within the next decade. We model the distribution of events as either a power law or lognormal distribution and use (1) Kolmogorov-Smirnov statistic to estimate goodness of fit, (2) bootstrapping to quantify the uncertainty in the estimates, and (3) likelihood ratio tests to assess whether one distribution is preferred over another. Our best estimate for the probability of another extreme geomagnetic event comparable to the Carrington event occurring within the next 10 years is 10.3% 95% confidence interval (CI) [0.9,18.7] for a power law distribution but only 3.0% 95% CI [0.6,9.0] for a lognormal distribution. However, our results depend crucially on (1) how we define an extreme event, (2) the statistical model used to describe how the events are distributed in intensity, (3) the techniques used to infer the model parameters, and (4) the data and duration used for the analysis. We test a major assumption that the data represent time stationary processes and discuss the implications. If the current trends persist, suggesting that we are entering a period of lower activity, our forecasts may represent upper limits rather than best estimates.

1. Introduction

Extreme space weather events are driven by the eruption of coronal mass ejections (CMEs). They have speeds, energies, and magnetic field strengths that are unusually large and can produce severe geomagnetic consequences [e.g., *Baker et al.*, 2008; *Cannon et al.*, 2013]. It has been suggested that many phenomena associated with such events follow a power law or quasi power law distribution [e.g., *Riley*, 2012], in terms of the frequency of events versus their severity. Such distributions are appealing because of their mathematical tractability but concerning because of their disproportionately large frequency, as compared with, say, Gaussian distributions. It remains to be established, however, to what extent these data really obey a power law distribution.

The "Carrington" event, named after Richard C. Carrington, who first observed and wrote about it, has risen to become the quintessential extreme space weather event for both what was and what was not known about it [*Carrington*, 1859]. Occurring just over 150 years ago, it was sufficiently recent that we have direct scientific links back to it but also far enough back that its description is far from complete. Until recently, the 1859 storm held the record as the largest space weather event observed in over 400 years [*McCracken et al.*, 2001]. It is likely that the Carrington CME was similar to the more recently studied 23 July 2012 event, which was directly observed in situ by the STEREO-A spacecraft [*Russell et al.*, 2013; *Baker et al.*, 2013; *Liu et al.*, 2014; *Riley et al.*, 2016].

Several previous studies addressed the likelihood of extreme space weather events. Under the assumption that four illustrative space weather data sets (flare intensity, coronal mass ejection speeds, *Dst*, and >30 MeV proton fluences) could be approximated by power law distributions, *Riley* [2012] estimated the likelihood of another Carrington-like event over the next decade. He found that the answer depended sensitively on (1) which parameter was used to define the extreme event and (2) what degree of severity qualified as an extreme event, concluding that the probability of another Carrington event, as defined by $Dst < -850$ nT, was approximately ~12% over the next decade. *Love* [2012] considered the 10 year occurrence probabilities of rare natural events, including geomagnetic storms, but added the important component of estimating uncertainties, using Poisson parameters to infer confidence-credibility intervals. He found, for example,

that the likelihood of a storm exceeding −589 nT was 17.8%, 68% CI [9.4, 27.8], or 95% CI [3.4, 38.6]. On one hand, his estimates matched well with those of *Riley* [2012], but the uncertainties, even at the 68% level, were huge, effectively covering the entire spectrum from negligible to reasonably likely. More recently, *Love et al.* [2015] considered the assumption that the data could be adequately described by a power law distribution. Specifically, they tested the idea of whether *Dst* would be better described by a lognormal distribution rather than a power law distribution. For the former, they estimated that a Carrington event ($Dst < -850$ nT) should occur about 1.13 times per century 95% CI [0.42, 2.41].

In this study, we extend these earlier studies in several important ways. First, we include a representative set of possible distributions in our analysis. Second, we remove the subjectivity associated with the choice of setting the minimum value of the tail component of the distribution. Third, we use bootstrapping to generate estimates of the confidence intervals. Fourth, we apply a likelihood ratio test for model selection. We focus on an analysis of *Dst*, deferring a more general study of other indicators of extreme events for the future.

2. Methods

2.1. Data

In a previous study, we chose four space physics data sets from which to assess the likelihood of an extreme event: peak rates from solar flares; the speeds of CMEs; the strength of geomagnetic storms, as determined from *Dst*; and >30 MeV proton fluences, which were, at the time, thought to be a reliable proxy for large solar energetic particles [*Riley*, 2012]. Here we focus on an analysis of *Dst*, for several reasons. First, the "events" contained within it can be directly associated with the main phases of magnetic storms [*Gonzalez et al.*, 1994]. Second, it is one of the most relevant parameters to consider from a societal impact perspective. Although subjective, we note that the 13 March 1989 geomagnetic storm, with a peak $Dst \sim -589$ nT, triggered a breakdown of the Hydro-Québec power grid and the consequential loss of electricity to six million people [*Bolduc*, 2002], as well as the meltdown of a transformer at the Salem, NJ, Nuclear Power Plant. On the other hand, the so-called "Bastille Day" event of 14 July 2000, which was associated with a peak *Dst* of −300 nT, had no notable effect on the power grid. Third, it displays a clear quasi power law profile, making it amenable to this type of analysis. Fourth, the data set spans almost 60 years, providing a long baseline for analysis.

We obtained final and preliminary *Dst* measurements from Kyoto [*Kyoto*, 2016], NOAA [*National Oceanic and Atmospheric Administration*, 2016], and NASA's OMNIWeb [*OMNIWeb*, 2016]. There are some minor differences amongst these data. For example, OMNIWeb's data run from 1964 to the present, relying on Kyoto's preliminary measurements for the most recent records. Both Kyoto's and NOAA's data sets begin in 1957, and while almost identical prior to 2008, they show deviations, particularly in the maximum extent of geomagnetic storms.

Generally speaking, a geomagnetic storm is a disturbance in the Earth's magnetosphere caused by variations in the solar wind. CMEs typically drive the largest storms, although high-speed streams can produce significant and often recurrent geomagnetic effects. Different phenomena in the solar wind result in different effects in the magnetosphere. For example, fast-mode shocks and their associated sheaths compress the magnetosphere, while extended periods of southward interplanetary magnetic field can transfer energy from the heliosphere to the magnetosphere [*Kivelson and Russell*, 1995]. A "storm" can be defined by *Dst* exceeding some negative value, such as −50 nT. More specifically, we can categorize storms as moderate (-50 nT> $Dst > -100$ nT), intense (-100 nT> $Dst > -250$ nT), or severe (-250 nT> $Dst > -600$ nT) [e.g., *Gonzalez et al.*, 1994]. Originally, the Carrington event of 1859 was thought to have had a peak negative *Dst* of −1760 nT [*Lakhina et al.*, 2005]. Later, however, this was adjusted to −850 nT [*Siscoe et al.*, 2006], a reduction by a factor of 2.

To generate an event-based data set, we define a "significant" magnetic storm as one in which $|Dst|$ exceeds 100 nT. Then, following *Love et al.* [2015], we identify all contiguous data for which this threshold is exceeded and label that as a single storm. Since it is also possible that *Dst* may cross the −100 nT threshold one or more times during the same storm; we define a further threshold of −20 nT to identify the start and end times of the storm. In practice, for computational efficiency, we employed a run length encoding algorithm that identified all contiguous intervals where $Dst < -20$ nT [*Robinson and Cherry*, 1967]. For each of these ∼11,000 events, we found the peak negative value for *Dst* (as well as its associated time). We then further pared down this list by only retaining those events that exceeded −100 nT, which reduced the total number of severe storms to 367.

2.2. Models

Here we outline a procedure for estimating the probability of a severe or extreme solar event, together with the uncertainties associated with that estimate [e.g., *Clauset et al.*, 2013]. In essence, we perform the following tasks: (1) Identify a set of possible distributions, which could reasonably describe the tail of the distribution; (2) Estimate the best model parameters; (3) Generate an ensemble of possible realizations using nonparametric bootstrapping, to compute the uncertainty associated with a particular forecast; (4) Estimate the significance (i.e., *p* values) that the data could be described by a power law; and (5) Compare other distribution models against the power law distribution.

2.2.1. Tail Distribution Models

We previously showed that a range of space weather phenomena are at least qualitatively consistent with a quasi power law distribution; that is, they demonstrate some degree of extended tail, beyond that which can be described by a Gaussian distribution [*Riley*, 2012]. Thus, we consider the following two types of distributions: power law (PL) and lognormal (LN). To this, we can add so-called "cutoff" distributions where the data dramatically drop off at some point.

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 lognormal 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. These two distributions, we will demonstrate, reasonably encompass the relevant phase space.

2.2.2. 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]. In particular, 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}:

$$P(x \geq x_{crit}) = \int_{x_{crit}}^{\infty} p(x')\mathrm{d}x' \tag{3}$$

which, for a finite data set, simplifies to

$$P(x \geq x_{crit}) = \frac{C}{\alpha - 1} x_{crit}^{-\alpha+1} \tag{4}$$

Hence, 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

$$\alpha - 1 = N \left[\sum_{i=1}^{N} ln \frac{x_i}{x_{min}} \right]^{-1} \tag{5}$$

where x_i are the measured values of x, N is the number of events in the data set, 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 equation (4), we can estimate the number of events as large as or larger than x_{crit} during the period covered by the data set:

$$E(x \geq x_{crit}) = NP(x \geq x_{crit}) \tag{6}$$

where N is the total number of events within the data set.

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 sometime Δt:

$$P(x \geq x_{crit}, t = \Delta t) = 1 - e^{-N\frac{\Delta t}{\tau}P(x \geq x_{crit})} \quad (7)$$

where τ is the total time span of the data set. Equations (4), (5), and (7) thus provide a robust technique for calculating the probability that an event of severity exceeding x_{crit} will occur sometime within the next Δt years.

Similar expressions can be written for the lognormal distribution, and equation (7) can be used to estimate probabilities based on this distribution.

2.2.3. 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 cutoff 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 [2012] and *Love et al.* [2015] identified the minimum values in severity (x_{min}) manually and, arguably, somewhat subjectively. In particular, *Riley* [2012] chose $x_{min} = 120$, while *Love et al.* [2015] used $x_{min} = 63$ as a lower bound for $|Dst|$. 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 [*Clauset et al.*, 2009]. This is defined by

$$D = \max_{x \geq x_{min}} |s(x) - P(x)| \quad (8)$$

where $s(x)$ is the CCDF of the data and $P(x)$ is the CCDF for the power law best fitting the data, both for $x \geq x_{min}$. The best estimate for x_{min} is the one that minimizes D, which can be seen as a balance between including more low-severity data to improve sample statistics and omitting data that may not reflect the true nature of the tail.

2.2.4. 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 are randomly sampled, and new pseudo data sets are constructed from the events drawn (and replaced). To each of these, one of the two distribution profiles (PL and LN) 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 data sets (i.e., bootstrap resamples) 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.

2.2.5. 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 the fact that the p value is large does not tell us which model matches the data most closely; 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 [*Joyce*, 2011]. The criterion 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 $p_{crit} = 0.05$.

3. Results

In Figure 1, we present a time series of these significant 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 [e.g., *Kilpua et al.*, 2015]. 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 through 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.

Figure 2 summarizes the probability estimates using the two possible distributions. Figure 2a shows the CCDF, which, as discussed above, is the probability that an event as large as or larger than some critical value will occur during a unit time interval. The open circles show all of 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*, 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 bootstrap resamples. Specifically, for each of 1000 bootstrapped pseudo data sets, a PL and LN 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, and (2) the LN profiles underestimate the low-severity events but capture the trends at higher severity.

Figure 2b summarizes the likelihood of observing an event as severe as or more severe than the most severe event observed, that is, $Dst < -589$ nT, during the entire span of the data (∼57 years). The probabilities were calculated for each bootstrap iteration and the distributions derived from the probabilities for each of them. For PL and LN distributions, the median probabilities are 0.95 and 0.63, respectively.

Using equation (7), 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 or lognormal 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 estimate is 20.3% 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 drop by almost a factor of 2.

Figure 3 summarizes the main statistical parameters for the power law 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 bootstrap resamples. Thus, as the number of bootstrap resamples is increased, our estimate for the different parameters improves. The best fit values are $x_{min} \sim 123$ 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. Following *Clauset et al.* [2009], and using the R package "PoweRlaw," we generated a large number of true power law synthetic data sets using a simple Monte Carlo procedure with the same parameters as those inferred from the analysis of the observed data. For each synthetic data set, we then fit a power law model and estimate the KS statistic. A p value was then constructed counting the fraction of the synthetic data sets for which the KS statistic was larger than that calculated for the empirical data. This is shown in Figure 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 Dst, ∼0.2, suggests that we cannot rule out the power law as the underlying distribution. On the other hand, this result

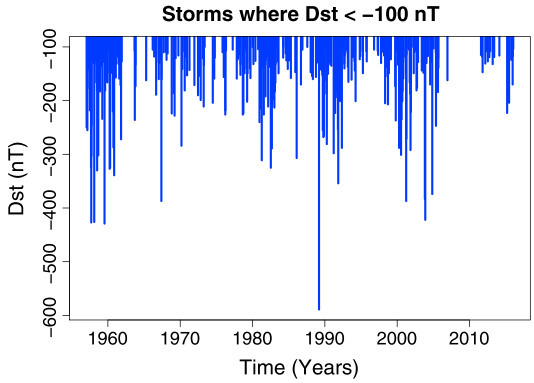

Figure 1. Magnetic storms, defined by events where |*Dst*| exceeds −100 nT, are shown as a function of time. Individual storms are identified as contiguous intervals where |*Dst*| exceeded −100 nT. The data were obtained from NASA's COHOWEB.

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.

We computed Vuong's test statistic (R_V) for the power law versus lognormal distributions: $R_V = -0.488$. Additionally, we calculated p values for both the one-sided and two-sided tests: 0.313 and 0.626, respectively. Larger positive values of R_V provide support for the first model over the second model. Thus, we can infer from these results that the lognormal distribution is slightly favored against the power law. However, for these

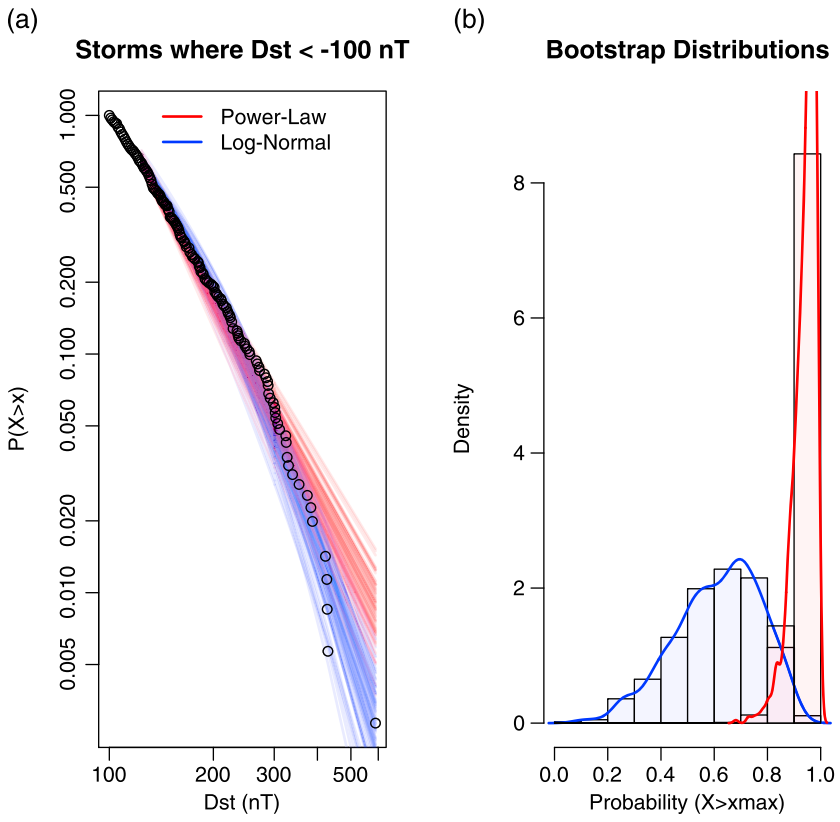

Figure 2. (a) Complementary cumulative distribution function (CCDF) for the geomagnetic storms shown in Figure 1. Bootstrap fits for the two distributions (power law and lognormal) are superimposed. (b) Histogram and density plots showing the probability of an event as large as or larger than the largest event in the data set (−589 nT) over the duration of the data set (~49 years). The density curve colors follow the convention given in the legend within Figure 2a.

Table 1. Best Estimates and Confidence Intervals for 10 Year Probabilistic Forecasts of |*Dst*| Exceeding −850 nT Assuming Power Law and Lognormal Distributions

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
Lognormal	3.0	0.6	9.0

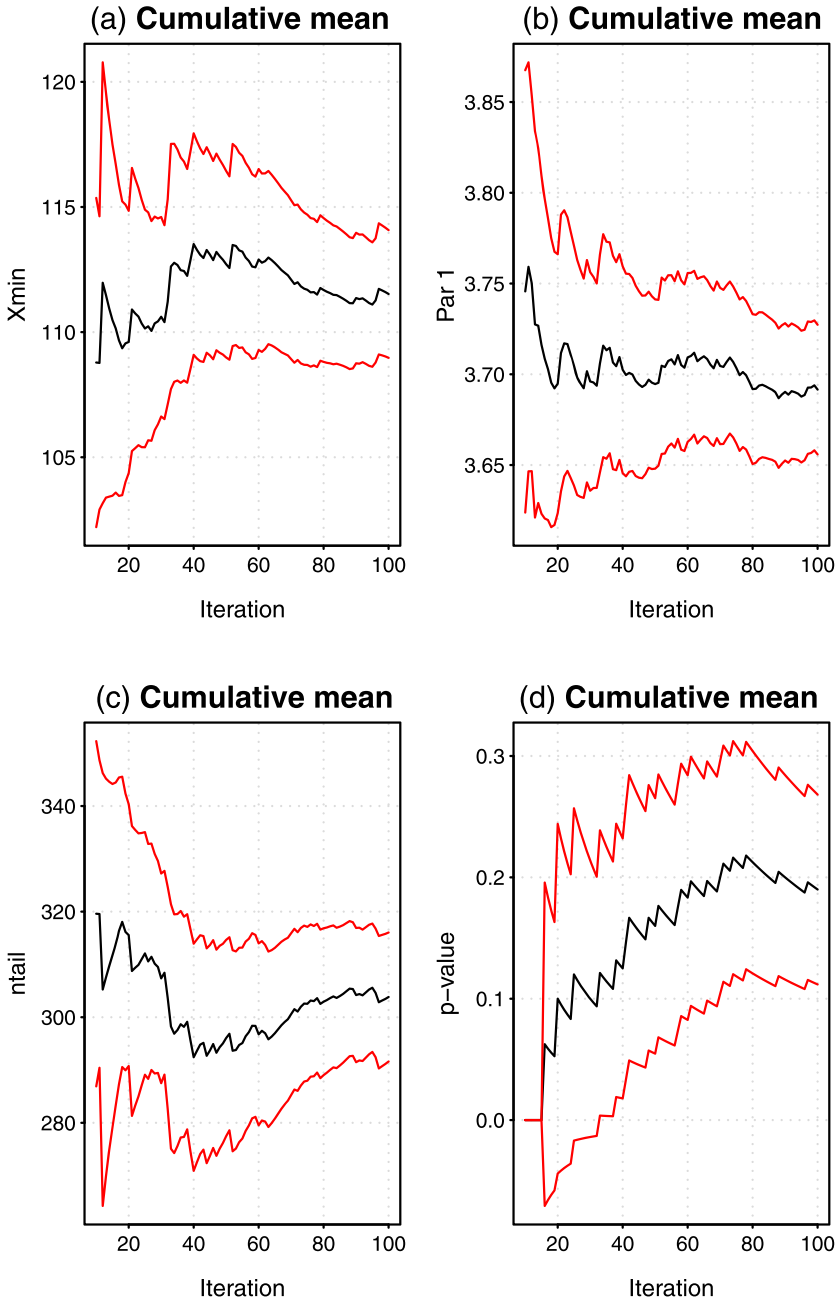

Figure 3. Summary of statistical parameters for power law bootstrap fit to the *Dst* data set as a function iteration, i.e., the number of bootstrap resamples (100 shown). The cumulative means of (a) x_{min}, (b) α (Par1), (c) n_{tail}, and (d) the p value.

Table 2. Best Estimates and Confidence Intervals for *Dst* for Each Solar Cycle From 1957 Through Early 2016

Cycle	Interval	Power Law	Lognormal
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]

results to be statistically significant, we require p values <0.05. Since the two p values exceeded this, we cannot discount either model.

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. Since the number of events following the end of cycle 23 was so small, we grouped cycle 24 events with cycle 23.

Table 2 summarizes the probabilities estimated using both the power law and lognormal 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. Probability Estimates for Solar Minimum and Solar Maximum

Our analysis thus far has been limited to forecasts that are averaged over a solar cycle or longer. However, it is clear from Figure 1 that there are strong variations in the frequency and magnitude of large geomagnetic storms on the scale of a solar rotation. Given that windows of 10–11 years are apparently not sufficient to compute accurate estimates of probabilities (section 3.1), we adopt a more limited analysis here to contrast forecasts for near-solar minimum and near-solar maximum conditions. Specifically, we extract data ± 2.5 years around each solar minimum and each solar maximum and combine them into two new data sets, one representing solar minimum-like and the other solar maximum-like conditions. This relies on the assumption that the events are independent of one another. A new effective time span is used for computing probability estimates ($2 \times 2.5 \times N_{min} = 2 \times 2.5 \times N_{max} = 30$ years), where N_{min} and N_{max} are the number of solar minima and maxima, respectively, contained within the data set.

Figure 4a shows the *Dst* time series with the solar minimum (green) and solar maximum (red) intervals marked. We note that although 2×2.5 years is less than half of the canonical 11 year solar cycle, because there is some variability in the duration of the cycle, and because of the asymmetry between the shorter rise and longer decay of the cycle, this means that some of the intervals overlap marginally (brown intervals). Nevertheless, this was necessary to produce data sets that were sufficiently populated with events and, equally, not omit too many events.

Figures 4b and 4d show the probability estimates (as discussed earlier with respect to Figure 2) for solar minimum (b) and solar maximum (d) conditions. We note that while the general shape of the profiles are the same (being bracketed by the lognormal and power law bootstrapped fits), there is a stark difference in the quantitative predictions (note the difference in scales on the x axes). Not only are there far fewer events during near-solar minimum conditions, but the average strength of events during these intervals is substantially less.

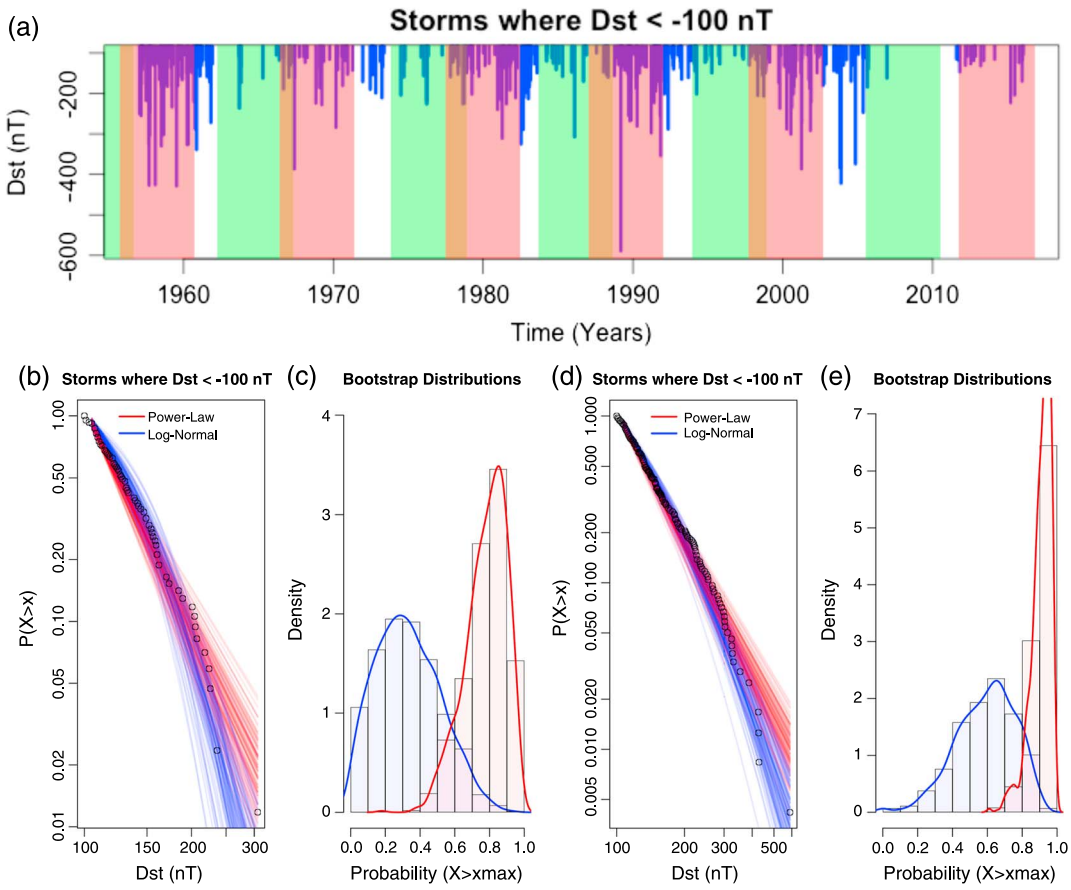

Figure 4. Comparison of probabilities for near-solar minimum and near-solar maximum conditions. (a) Time series of $Dst < -100$ nT with boxes overlaid bracketing (±2.5 years) solar minimum (green) and solar maximum (red) conditions. (b and c) As in Figure 2 for solar minimum conditions. (d and e) As in Figure 2 for solar maximum conditions. Note the different *x* axis limits in Figures 4b and 4d.

Figures 4c and 4e similarly show the likelihood of observing an event as severe as or more severe than the most severe event observed in that data set. Thus, while the power law distributions suggest a strong likelihood of observing such an event in both cases, the largest event in the solar minimum data set was just over −300 nT, while the largest solar maximum event was almost −600 nT. A second distinction of note is that the lognormal distribution moves closer to the power law distribution during solar maximum conditions, perhaps suggesting that forecasts at solar maximum are not as sensitive to the assumed distribution.

By computing forecasts for each of these data sets, we estimate that the median likelihood of another event as large as or larger than −850 nT, assuming a power law distribution, is 1.4% during solar minimum conditions and 28% for solar maximum conditions.

4. Summary and Discussion

In this report, we have described and applied a general technique for assessing both the likelihood and uncertainties of an extreme space weather event. We considered both power law and lognormal distributions as alternatives for explaining the observed distribution in severity. We inferred that the probability of another event within the next decade exceeding −850 nT was 20.3% 95% CI [12.5, 30.1] for a PL distribution but only 3.0% 95% CI [0.6, 9.0] for a LN distribution.

Although the analysis presented here incorporated more sophisticated statistical methodologies than our previous studies [*Riley*, 2012; *Love et al.*, 2015], somewhat paradoxically, our analysis produced probabilistic forecasts that are even less constrained. Because we cannot be confident of the underlying distribution and, indeed, it appears that the data lie firmly between the power law and lognormal distributions, our basic estimate could be as low as 3.0% or as high as 20.3%. 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 timescale 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 data set 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 some that the most extreme storms tend to occur at or near solar minimum.

Similarly, over longer timescales, there is ample evidence for nonstationarity [e.g., *Riley*, 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 our analysis of decadal subsets of *Dst* did not yield any systematic trend in forecasts for Carrington events, constructing two long-window data sets (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 1950's/early 1960's was replaced with the unusually quiet period surrounding and following the 2008/2009 solar minimum (aka the "Eddy" minimum). 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 year) data set.

The second major assumption addressed in this study is whether the data are better represented by a PL or LN distribution. We inferred that both the LN and PL distributions were consistent with the *Dst* data set. 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 [*Angelopoulos et al.*, 1999; *Klimas et al.*, 2000], which provides a natural explanation for the presence of power law distributions. By extension, we could posit that in analogy with the Abelian Sandpile model, the magnetotail becomes progressively loaded until some specific threshold is reached and then reconnects and produces the observed *Dst* maxima. However, this is, undoubtedly, a simplistic interpretation of a considerably more complex system. *Love et al.* [2015] have 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 lognormal 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, since the errors would not be expected to be distributed normally in log-log space. Second, in any finite-sized system, there must be a cutoff at some point. The key issue is whether that cutoff 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 cutoff 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 *Dst*, 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 $\mathbf{J}_{\perp} \sim 2500(\mathbf{B} \times \nabla)\mathbf{B}/B^2$, 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 cutoff 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 as or larger than 2500 nT is ~1.3%. Thus, our estimate of 20.6% would only be reduced to ~19.3%.

An interesting but as yet unexplored possibility is that if we could provide firm limits to the cutoff and, additionally, successfully argue for a lognormal distribution, this would allow us to set the rightmost portion of the curve, allowing 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%, which is larger than two earlier estimates of ~12% [*Riley*, 2012] and ~11% [*Love et al.*, 2015], although certainly within the overlapping confidence intervals. The disagreement between the current value and that in *Riley* [2012] is due, primarily, to the addition of data from 1957 through 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 lognormal distribution. In particular, the incorporation of the low-severity events strongly influenced the fit of the lognormal 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 data set under study as well as the techniques used to analyze them. Additionally, it is worth noting that it is not only the forecast estimates that must be communicated but also the uncertainties and assumptions that accompany them *Love* [2012].

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 23 July 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 et al.*, 2008]. Our estimate for an extreme space weather event is 50%/decade (PL distribution) or 10% (LN distribution), which are 12 and 2.5 times larger than the earthquake forecast. However, even these comparisons can be misleading because it is the consequences of each disaster that society cares more about.

Ironically, in this study, we set out to firmly establish whether a power law or lognormal distribution better fits the data. However, 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 data sets, which intervals, and which distributions were used to make the estimates, the results varied substantially. 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, and (4) a good definition of *Dst* for an extreme event is that $Dst < -850$ nT, we conclude that our best estimate for the probability of such an event over the next decade is approximately 10% 95% CI [1, 20].

Acknowledgments
The authors gratefully acknowledge the support of NSF's Frontiers in Earth System Dynamics (FESD) and NASA's Living with a Star (LWS) program, under which this work was performed. The *Dst* index was obtained from the Kyoto World Data Center (https://wdc.kugi.kyoto-u.ac.jp/dstdir/).

References

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

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

Baker, D. N., et al. (2008), *Severe Space Weather Events—Understanding Societal and Economic Impacts*, Nat. Acad. Press.

Bell, J. T., M. S. Gussenhoven, and E. G. Mullen (1997), Super storms, *J. Geophys. Res.*, *102*, 14,189–14,198, doi:10.1029/96JA03759.

Bolduc, L. (2002), GIC observations and studies in the Hydro-Québec power system, *J. Atmos. Sol. Terr. Phys.*, *64*, 1793–1802, doi:10.1016/S1364-6826(02)00128-1.

Cannon, P., et al. (2013), *Extreme Space Weather: Impacts on Engineered Systems and Infrastructure*, Royal Academy of Engineering, London.

Carrington, R. C. (1859), Description of a singular appearance seen in the Sun on September 1, 1859, *Mon. Not. R. Astron. Soc.*, *20*, 13–15.

Clauset, A., C. Shalizi, and M. Newman (2009), Power-law distributions in empirical data, *SIAM Rev.*, *51*(4), 661–703, doi:10.1137/070710111.

Clauset, A., et al. (2013), Estimating the historical and future probabilities of large terrorist events, *Ann. Appl. Stat.*, *7*(4), 1838–1865.
Dessler, A. J., and E. N. Parker (1959), Hydromagnetic theory of geomagnetic storms, *J. Geophys. Res.*, *64*(12), 2239–2252.
Efron, B., and R. J. Tibshirani (1994), *An Introduction to the Bootstrap*, CRC Press, Boca Raton, Fla.
Field, E., K. Milner, and 2007 Working Group on California Earthquake Probabilities (2008), Forecasting California's earthquakes—What can we expect in the next 30 years?, USGS Fact Sheet 2008–3027, U.S. Geol. Surv., California.
Gonzalez, W. D., J. A. Joselyn, Y. Kamide, H. W. Kroehl, G. Rostoker, B. T. Tsurutani, and V. M. Vasyliunas (1994), What is a geomagnetic storm?, *J. Geophys. Res.*, *99*, 5771–5792, doi:10.1029/93JA02867.
Joyce, J. M. (2011), Kullback-Leibler divergence, in *International Encyclopedia of Statistical Science*, pp. 720–722, Springer, New York.
Kilpua, E., N. Olspert, A. Grigorievskiy, M. Käpylä, E. Tanskanen, H. Miyahara, R. Kataoka, J. Pelt, and Y. Liu (2015), Statistical study of strong and extreme geomagnetic disturbances and solar cycle characteristics, *Astrophys. J.*, *806*(2), 272.
Kivelson, M. G., and C. T. Russell (1995), *Introduction to Space Physics*, Cambridge Univ. Press, Cambridge.
Klimas, A. J., J. Valdivia, D. Vassiliadis, D. Baker, M. Hesse, and J. Takalo (2000), Self-organized criticality in the substorm phenomenon and its relation to localized reconnection in the magnetospheric plasma sheet, *J. Geophys. Res.*, *105*(A8), 18,765–18,780.
Kyoto (2016), Geomagnetic Equatorial *Dst* Index Home Page. [Available at: https://wdc.kugi.kyoto-u.ac.jp/dstdir/.]
Lakhina, G. S., S. Alex, B. T. Tsurutani, and W. D. Gonzalez (2005), Research on historical records of geomagnetic storms, in *Coronal and Stellar Mass Ejections*, IAU Symposium, vol. 226, edited by K. Dere, J. Wang, and Y. Yan, pp. 3–15, Cambridge Univ. Press, Cambridge., doi:10.1017/S1743921305000074
Liu, Y. D., et al. (2014), *Observations of an Axtreme Storm in Interplanetary Space Caused by Successive Coronal Mass Ejections*, vol. 5, 3481.
Lockwood, M., A. P. Rouillard, and I. D. Finch (2009), The rise and fall of open solar flux during the current grand solar maximum, *Astrophys. J.*, *700*, 937–944, doi: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*, L10301, doi:10.1029/2012GL051431.
Love, J. J., E. J. Rigler, A. Pulkkinen, and P. Riley (2015), On the lognormality of historical magnetic storm intensity statistics: Implications for extreme-event probabilities, *Geophys. Res. Lett.*, *42*(16), 6544–6553.
McCracken, K. G., G. A. M. Dreschhoff, E. J. Zeller, D. F. Smart, and M. A. Shea (2001), Solar cosmic ray events for the period 1561–1994: 1. Identification in polar ice, 1561–1950, *J. Geophys. Res.*, *106*, 21,585–21,598.
McMorrow, D. (2009), Rare events, *Tech. Rep.*
Newman, M. (2005), Power laws, Pareto distributions and zipf's law, *Contemp. Phys.*, *46*, 323–351.
National Oceanic and Atmospheric Administration (2016), The Disturbance Storm Time Index. [Available at: https://www.ngdc.noaa.gov/stp/geomag/dst.html.]
OMNIWeb (2016), OMNIWeb. [Available at: http://omniweb.gsfc.nasa.gov/form/dx1.html.]
Riley, P. (2012), On the probability of occurrence of extreme space weather events, *The Int. J. Res. Appl.*, *10*, S02012.
Riley, P., J. A. Linker, Z. Mikic, and R. Lionello (2000), Solar cycle variations and the large-scale structure of the heliosphere: MHD simulations, in *The Sun and Space Weather, 24th Meeting of the IAU, Joint Discussion*, vol. 7, Manchester, U. K.
Riley, P., C. Schatzman, H. V. Cane, I. G. Richardson, and N. Gopalswamy (2006), On the rates of coronal mass ejections: Remote solar and in situ observations, *Astrophys. J.*, *647*, 648–653, doi:10.1086/505383.
Riley, P., R. M. Caplan, J. Giacalone, D. Lario, and Y. Liu (2016), Properties of the fast forward shock driven by the July 23, 2012 extreme coronal mass ejection, *Astrophys. J.*, *819*, 57, doi:10.3847/0004-637X/819/1/57.
Riley, P., 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.
Robinson, A., and C. Cherry (1967), Results of a prototype television bandwidth compression scheme, *Proc. IEEE*, *55*(3), 356–364.
Russell, C. T., et al. (2013), The very unusual interplanetary coronal mass ejection of 2012 July 23: A blast wave mediated by solar energetic particles, *Astrophys. J.*, *770*, 38, doi:10.1088/0004-637X/770/1/38.
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.
Siscoe, G., N. U. Crooker, and C. R. Clauer (2006), *Dst* of the Carrington storm of 1859, *Adv. Space Res.*, *38*, 173–179, doi:10.1016/j.asr.2005.02.102.
Vasyliūnas, V. M (2011), The largest imaginable magnetic storm, *J. Atmos. Sol. Terr. Phys.*, *73*(11), 1444–1446.
Vuong, Q. H. (1989), Likelihood ratio tests for model selection and non-nested hypotheses, *J. Econometric Soc.*, *57*(2), 307–333.