A second look at “On the probability of occurrence of extreme space weather events”, by P. Riley[1]

analysis by Stephen Parrott

1 Introduction

P. Riley’s “On the probability of occurrence of extreme space weather events” [1] has been widely quoted in the popular press. It gives the impression that from accepted statistical analysis one can predict that the probability of a solar storm worse than the worst ever observed (the Carrington event) in the next decade is on the order of 12%, surprisingly high. This estimate was obtained from a power law model. The conclusion of [1] speaks for itself:

> “In closing, we reiterate that our primary aim in this study was to introduce a technique for estimating the probability of occurrence of extreme space weather events. Additionally, our analysis has shown that a relatively rich subset of space physics data can be approximated by power law distributions. Our results allowed us to answer a basic question, at least in an approximate way: How likely are such events? . . . our results overall suggest that the likelihood of another Carrington event occurring within the next decade is $\sim 12\%$.”

Even a cursory reading of Riley’s paper will show how shaky is this conclusion. It obtains probability estimates from various data sets in various ways unified by a power law assumption. The estimates vary over two orders of magnitude, from 1.5% (paragraph [39]) to an unbelievable 85% (paragraph [31]).

The paper itself rejects the 85% estimate as “not credible” (paragraph [31]). However, it was obtained by the same methods which yielded the 12% estimate using data equally reliable. The 85% estimate is obtained by applying the power law assumption to data probably comprising over a thouand observations of Coronal Mass Ejections (CME). After this estimate turned out too high to be credible, the paper reapplied the same method to a subset of a few tens of the CME’s to obtain the 12% estimate (which was also obtained from another data

[1] P. Riley, *Space Weather* **10** (2012), S02012

set). Such apparent "cherry-picking"[2] of the data needs to be carefully justified. The paper does briefly address this issue, but unconvincingly in my opinion.

A previous analysis [3] of Riley's paper made this and more technical points. There were two main criticisms.

One questioned the power law model itself. The other noted that the paper is so vaguely written as to make it difficult (in some cases) and impossible (in others) to check the arithmetic leading to its various probability estimates. However, sometimes it was possible to guess (e.g., from the figures) the assumptions sufficiently to check the arithmetic, and in these cases my arithmetic usually produced estimates differing by an order of magnitude from the paper's.

Since I posted [3], I have found new guesses which enable one to obtain some of the paper's probability estimates. These will be reported below.

I have also stumbled upon what seem surprising anomalies in the data reported. These will be discussed in detail.

Each data set has its own quirks and apparent anomalies, but they have a common core which is easy to perceive when pointed out, but which I find very puzzling. They seem almost like a detective mystery waiting for some Sherlock Holmes.

2 Introduction to the mystery

This section concisely introduces the mystery for those who already have some familiarity with Riley's paper [1] . Full discussion and definitions will be postponed to subsequent sections.

Riley's paper analyzes four data sets, Solar Flares, Coronal Mass Ejections (CME), Geomagnetic Storms, and Ice Core Samples. For each data set there is a Figure(a) and a Figure(b), e.g., Figures 2(a) and 2(b) for Solar Flares.

Figure (a) is a histogram constructed from observations. Roughly speaking, it may be considered as representing the log-log graph of the probability density function (pdf) $p = p(x)$ for the probability distribution from which the data were drawn,[3] which is assumed to be a power law:

$$p(x) = \begin{cases} C_\alpha x^{-\alpha} & x_{min} < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Here $\alpha > 1$ and $C_\alpha := (\alpha - 1)x_{min}^{\alpha-1}$ is the constant which normalizes the total probability to 1.

The paper's equation (1) which defines the power law does not mention x_{min} (and is therefore incorrect), but inclusion of x_{min} is necessary for the mathematics to make sense. One of the difficulties in interpreting the paper's analyses of the various data sets is that the value of x_{min} is never explicitly given, but has to be guessed from the figures or other given data.

[2] i.e., selective rejection of parts that do not support a desired conclusion

[3] More precisely, before passing to log-log coordinates, the histogram is an approximation to a constant times graph of the pdf, i.e. the histogram is not normalized to total probability 1.

Figure (b) is a graph, not a histogram. It is the experimentally observed graph, in log-log coordinates, of the complementary cumulative distribution function (CCDF) P, which for a pdf $p = p(x)$, Riley denotes[4]

$$P(x \geq x_{crit}) = \int_{x_{crit}}^{\infty} p(x')\, dx' \quad . \qquad \text{Riley's (2)}$$

When p is given by the power law $p(x) = C_\alpha x^{-\alpha}$, which is the paper's premise, then the CCDF is also given by a power law but with exponent $\alpha - 1$ one less than the exponent α of the pdf:

$$P(x \geq z) = \left(\frac{z}{x_{min}} \right)^{-(\alpha-1)} \quad .$$

Here is the mystery. Since Fig. (a) is in log-log coordinates, it should approximate a straight line with slope $-\alpha$. Fig. (b) should approximate a line with slope $-(\alpha - 1)$. But for three of the four data sets, the slopes of Figures (a) and (b) are almost the same!

The exception is the Geomagnetic Storm data set, for which the slope of Fig. (a) is -3.7, compared to -3.0 for Fig. (b). That is still a slope quite a bit steeper for Fig. (b) compared to the expected -2.7.

Table 1 gives all the measured slopes, accurate to at least ± 0.1. "Accuracy" refers to estimated maximum error, which is never more than ± 0.1—typical errors are probably much less. Appendix 2 on measurement explains how the accuracy is determined.

Riley's claimed slopes are described in syntax often making it ambiguous whether they refer to Fig. (a) or (b). The previous analysis [3] assumed that they referred to (a), but in light of subsequent information, I now would guess that at least some of them refer to (b).[5] When ambiguous, the claimed slope is listed under Fig. (b) even though it might refer to (a).

The measured "slope" of Fig. (a) refers to the dashed line of best fit according to the usual "least squares" (LS) method. Riley doesn't furnish a line of best fit with the "maximum likelihood estimate" (MLE) of slope for Fig. (a), which is strange given that most of the final probability estimates are calculated using that MLE estimate. If an MLE line had been furnished for Figs. (a), in most cases it would have been obvious that it didn't well fit the data and so should not have been used for the probability estimates. For Figs. (b), Riley furnishes both a dashed LS line and a solid MLE line.

[4] Riley's notation is unusual. In his $P(x \geq x_{crit})$, the lower-case x stands for a random variable, but x_{crit} is a real number. When quoting Riley or discussing his equations, we have to use his notation, but elsewhere our notation will be either explicitly defined or standard. In particular, when we write $p(x) = C_\alpha x^{-\alpha}$, we mean that p is the function which assigns to a real variable x the number $C_\alpha x^{-\alpha}$. It would probably have been better to use a different letter than x, since Riley uses x for a random variable, but this was only noticed after all the type was set.

[5] A message to the author specifically asking about this was ignored, along with three other messages asking about various points in [1].

TABLE 1 --- measured slopes

Figure	Measured slope	accuracy	Riley's claimed "slope"
2(a) dashed LS	-0.84	+-.09	-1.8 (unambiguous)
2(b) solid MLE	-0.84	+-.04	-0.84 (unambiguous)
2(b) dashed LS	-0.85	+-.04	
4(a) dashed LS	-3.11	+-.09	
4(b) solid MLE	-3.12	+-.06	-3.2 (ambiguous)
4(b) dashed LS	-3.29	+-.06	
5 solid MLE	-6.08	+-.02	-6.1 (ambiguous)
5 dashed LS	-6.63	+-.02	
8(a) dashed LS	-3.69	+-.07	
8(b) solid MLE	-3.00	+-.09	-3.2 (ambiguous)
8(b) dashed LS	-3.05	+-.09	
10(a) dashed LS	-1.84	+-.04	
10(b) solid MLE	-1.98	+-.02	-2.0 (ambiguous)
10(b) dashed LS	-2.09	+-.02	

3 History of this paper

Having presented the mystery on which the reader can ponder, it seems appropriate to explain how I came on it. Most readers would be unlikely to notice that the slopes of Figures (a) and (b) are generally inconsistent with the paper's power law premise, and I did overlook this for months. When my first analysis [3] was posted, I was unaware of it.

When I first read in the popular press Riley's extensively reported estimate that there is about a 12% probability of a magnetic storm worse than Carrington in the next decade, I was skeptical. An initial reading of his paper [1] did not reassure me.

His equation (1) defines "power law" as a pdf of the form $p(x) = Cx^{-\alpha}$ where "C is a constant determined from where the power law intercepts the y axis". This is incorrect. The graph of a power law *never* intercepts the y-axis. For $\alpha > 1$, it cannot approach the y-axis because then the integral of p over the positive x-axis would diverge. The paper's equation (1) definition of "power law" entirely omits the critical parameter x_{min} that defines the interval $[x_{min}, \infty)$ which is the domain of p. "How well could a paper with such a glaring error in its very first equation have been refereed?", I asked myself. A bad typo in its equation (7) did not help matters.

In view of that, I was motivated to check everything. I could see immediately that the paper's "cherry-picking" of the Coronal Mass Ejection (CME) data[6] was highly questionable, but I also wondered if the paper's probability estimates had been correctly obtained even assuming the legitimacy of the "cherry-picking". The discussion of the CME data is unacceptably vague. In particular, the paper's paragraph [24] reference to "a MLE fit" and "a revised slope of -6.1" is ambiguous as to whether this refers to the slope of the pdf or the CCDF.

I wrote the author, asking which was meant. He chose not to reply, nor did he reply to three other courteously worded inquiries about various points in the paper. Even at that early stage, I had noticed the anomaly that the slopes of Figures 4(a) and 4(b) seemed about the same, though I wasn't sure because I had not learned how to accurately measure the .pdf file of the paper.[7] I was measuring the original published figures with a physical ruler, which didn't afford enough resolution to be sure that their slopes could not differ by one as they should. I raised that point in one of my unanswered messages.

Careful authors of scientific papers are usually happy when anyone reads their work carefully enough to ask detailed questions. If an author refuses to reply, this is cause for reasonable suspicion that he may have no good reply.

[6] After an analysis of CME's above 700 km/sec produced a result which the paper characteriazed as "not credible", it restricts to a few tens of CME's above 2000 km/sec to obtain its main estimate that the probability of an event worse than Carrington in the next decade is 12%. If the paper's methods were valid, then the original analysis of speeds over 700 should have produced a believable result.

[7] Here ".pdf" refers not to "probability density function" abbreviated "pdf" but to a computer file with extension ".pdf", like "riley.pdf", which is intended for the Adobe Acrobat application.

I decided that the only way to resolve the ambiguity was to analyze independently the original raw data. On the website which Riley cites as furnishing his data for Coronal Mass Ejections (CME),[8] I found only 20 observations of CME's above 2000 satisfying the paper's stated criteria. These implied an exponent $\alpha = 5.9$ for a pdf $p(x) = Cx^{-\alpha}$, which seemed reasonably close to the paper's ambiguous claim of "revised slope of -6.1 " if "revised slope" referred to $-\alpha$.

The paper's analyses of the other data sets also used ambiguous language which did not clearly distinguish between slopes of pdf and CCDF. However, the language of the other sets was perhaps weighted slightly more toward the CCDF interpretation.

Since the author had already ignored four inquiries, there seemed no hope of obtaining clarification from that quarter. I made the assumption, in retrospect perhaps too hastily, that the same interpretation of phrases similar to "MLE slope" would apply to all data sets. I thought that the issue had already been settled in favor of α rather than $\alpha - 1$ for the CME data, and applied the same assumption to the other data sets. This resulted in probability estimates (for the probability of an event worse than Carrington in the next decade) which were generally too high to be believable. These were reported in my initial analysis [3] of Riley's paper. Later I noticed that if for some of the data sets, one assumed that "MLE fit" and the like referred to the CCDF instead of pdf, one obtained Riley's probability estimates (possibly under other auxiliary assumptions which seemed conceivable).

After David Roodman showed me how to measure the high-resolution .pdf of Riley's paper under high magnification, I noticed that the paper's Figure 5 showed many more than the 20 observations of CME's above 2000 that I had extracted from the raw data. In the printed version, the individual observations are so close together that they cannot be distinguished, but under high magnification of the .pdf file they can be. They are still a bit hard to distinguish, but they could be counted, though I have not done so. I estimate that there are about 50 observations in that range, definitely far above the mere 20 that I found in the original data. When 50 observations instead of 20 are used with the interpretation that the paper's "revised slope" of -6.1 refers to the slope $\alpha - 1$ instead of α, the paper's final probability estimate of 12% is obtained.

This calls into question my original assumption that "MLE fit" and the like would refer to the slope of the pdf, as was my previous best guess. My current best guess in light of the information which has surfaced in the meantime is that Riley's references to "MLE fit" and the like may refer to the slope of the CCDF. How to obtain the paper's probability estimates from this assumption will be presented in the discussions of the individual data sets.[9]

[8] Cited in paragraph [29] as /http://cdaw.gsfc.nasa.gov/CME_list/

[9] It's is not clear why Riley obtained so many more CME's with speeds above 2000 than I did. It may be due to data selection. He reports so-called "quadratic speeds" at "the highest measurement possible". That is also what I reported, so we should have obtained the same number. For some observations the website containing the data reports a "linear speed" but no "quadratic speed". (That happens when only two measurements were made.) If Riley

Had the author answered my initial query as to his meaning for "MLE fit" of the CME data, the present paper might never have been written. The previous analysis [3] would still have been posted but it might have been written differently, taking into account the author's response.

4 The Solar Flare data set–Part 1

The original "Analysis" [3] ignored this data set because it is the only one of the four which does not yield a final estimate for the probability of a solar storm worse than Carrington in the next decade. (That is because the strength of solar flares associated with Carrington is unknown.) However, recently I have found a major anomaly which is hard to understand. The following quote from the paper's paragraph [24] introduces it:

> "In Figure 2b we show the CCDF, i.e., the probability of an event occurring that exceeds some critical peak rate, as a function of peak rate. . . . We note that the slope of the event-frequency plot is ~ -1.8 in agreement with previous studies [*Lin et al., 1984; Dennis, 1985*], while the slope (computed using MLE) for the CCDF is -0.84. Theoretically, we would expect the latter to be one less than the former, which, given the errors associated with computing the former, is relatively consistent."

The quoted study of Dennis [4] (which uses a different data set than Riley) does claim power law behavior $p(x) = Cx^{-\alpha}$ with exponent $\alpha \sim 1.8$, and Lin, et al. [5] plot an observed CCDF with slope which they estimate as "about -1", which would give $\alpha = 1 + 1 = 2$. As Riley indicates, the difference between $\alpha = 1.8$ and $\alpha = 2$ is probably not significant in this context.

Strangely, Riley seems unaware that the measured slope of his Figure 2(a) is -0.8, as opposed to his paragraph [24] claim that it is -1.8.[10] [11] By "measured

defaulted to "linear" when "quadratic" was not available, that could explain the difference.

However, the large difference between his MLE estimate $\alpha = 7.1$ ($\alpha - 1 = 6.1$, assuming that's what he meant) and the MLE estimate $\alpha = 5.9$ obtained from my data is unexplained and noteworthy.

All else being equal, this difference can make a difference of an order of magnitude for the final probability estimate. In this case, the final estimates happened to be not much affected because of a difference in the number of observations used to calculate the MLE (which appears in the final estimate but not the MLE itself). (That is, "all else" was not equal.) But usually, one would expect the final result to be greatly affected. This gives yet another reason that Riley's estimates should be independently checked before being cited without caveats.

[10] Since the scales for Fig. 2(a) are different on the horizontal and vertical axes, I should make clear that "measured slope" means the slope that would be seen were the scales the same, i.e., "measured slope" should approximate $-\alpha$ for a pdf $p(x) = Cx^{-\alpha}$. See Appendix 2 for the details of how the measurements were carried out.

[11] Could the author have obtained an MLE estimate of slope -1.8 and simply assumed without checking that it would well represent the slope of Figure 2(a)? If so, that should be taken into account in assessing the probable accuracy of other statements, particularly MLE estimates in ambiguous contexts.

slope", I mean the slope of the dashed LS line in Fig. 2(a) which appears to well fit the data over the range $[10^4, 10^5]$ for which it was calculated. If a solid MLE line with slope -1.8 were plotted on Riley's Figure 2(a), it would obviously *not* well fit any portion of the data.

I cannot conveniently exhibit this in a figure because there are too many data points in Riley's Figure 2(a) to plot by hand. and I do not have the facilities to undertake an independent analysis of large quantities of raw data. However, the reader can verify this by obtaining a high-resolution .pdf copy of [1] from the Wiley website given in footnote 13 and using the "measuring tool" (a part of the Adobe Acrobat program) at 400% magnification to construct an MLE line of slope -1.8. The difference between the dashed LS line with slope -0.8 and the MLE line does not look enormous, but it is different enough that it could not be a product of some error. The portions of the data which do appear to lie more or less on a line of negative slope but are *not* in the range $[10^4, 10^5]$ have absolute slope even less than the 0.8 of the dashed line. There is no way that a line of slope -1.8 could fit this data.

This is not the only major anomaly in the paper's discussion of the solar flare data—the other will be discused in the Solar Flares–Part 2 section. Neither anomaly affects the paper's probability estimates, but does call into question its care in the data analysis. It is just another reason (starting with its incorrect equation (1)) to check everything independently before accepting the paper's conclusions.

5 The ice core data set

The paper's paragraph [40] introduces the ice core data:

> "[40] Finally in the chain of space weather parameters from the Sun to the Earth, we arrive at space weather records potentially contained within ice cores. The value of these data lies in their long time span going back more than 400 years; however, they are not without caveats. First, while the nitrate spikes are generally believed by space physicists to be a record of large, historical space weather events, *[McCracken, et al., 2001],* ice core chemists are skeptical. They posit that no viable mechanism exists by which Solar Proton Events 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."

Since Riley's paper was published, Wolff, et al., [7] has appeared which asserts the following:[12]

- A major nitrate spike in 1859 (the year of Carrington) is observed in only one out of five Greenland ice cores. The core in which it was observed is the one from which were taken the 70 observations of McCracken, et al.,

[12] I am indebted to David Roodman for this reference.

[2] which form Riley's data set. No spike in that year was observed in eight Antartic ice cores.

- The dating to 1859 of that major spike is questionable. Other Greenland ice cores show a spike in 1863 which may correspond to the 1859 spike of McCracken, et al.

- Detailed chemical analyses of the 1863 spike just mentioned (which was not undertaken for the McCracken, et al., spike) show signatures of nitrate from forest fires. This could be observed in Greenland, but not in distant Antartica.

- Wolff, et al. [7], concludes:

  > "In summary, the nitrate event identified as 1859 [by McCracken, et al., [2]] is most likely the same event that more recent Greenland ice cores identify at 1863. The parallel event in other cores, as well as all other significant nitrate spikes in those cores, has an unequivocal fingerprint of a biomass burning plume. . . . In any case, the [McCracken, et al.] core is the only one of the 8 Antartic and 6 Greenland cores . . . that claims a spike in 1859. Taking the data from all the cores discussed here, we can say clearly that an episode of the size of the Carrington Event has not left an observable imprint in nitrate in ice."

I am not qualified to undertake a scientific evaluation of these assertions, but personally, I find them convincing enough to reserve judgment of any conclusions based on the McCracken, et al., data until the claims are evaluated by experts. (Riley's analysis of the ice core data is based solely on the data of McCracken, et al.)

Although privately, I am skeptical of the ice core data, nevertheless a discussion of Riley's analysis will well illustrate the main mystery which this paper explores—that three of Riley's four data sets, Figures (a) and (b) have almost the same slope although these slopes should differ by 1 if the paper's power law assumption holds.. Since the McCracken data is so small, analysis by hand is feasible.

Our Figure 1(a) redraws Riley's Figure 10(a) histogram for the ice core data, using coordinates for the data bins measured from a high-resolution .pdf file of Riley's paper.[13] Subfigure 1(a)(i) includes Riley's dashed line of best least squares (LS) fit. The small circles represent the data bins. The left-most circle probably was not included by the author in the calculations of the dashed LS line, and the right-most point may have been omitted as well. [14]

[13] Obtainable at http://onlinelibrary.wiley.com/doi/10.1029/2011SW000734/abstract .

For comparison with Riley's paper, Figure 1(a)(i) attempts to replicate Riley's Figure 10(a) rendering of the data rather than to plot the pdf from the original data. There are some anomalies in Riley's Figure 10(a), but they would not affect the discussion.

[14] See the descriptions around Riley's Figs. 10 and 2 for the author's explanations. Our Figs. 1(a) omit for clarity the vertical dashed lines in Riley's Fig. 10(a) which are connected

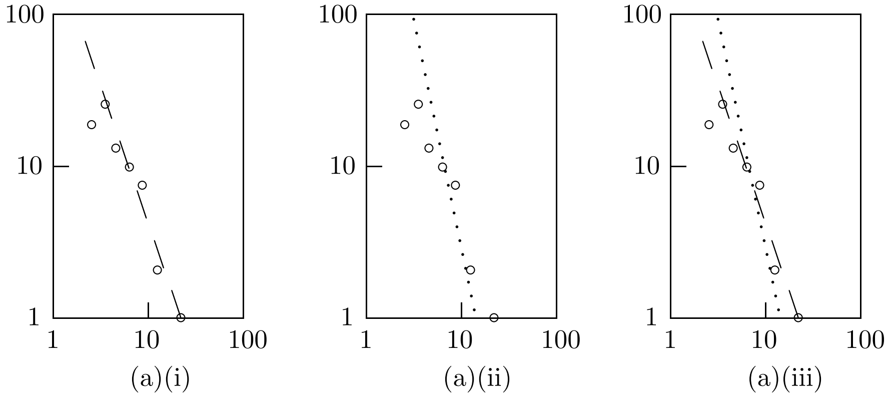

Figure 1: Subfigure 1(a)(i) redraws Figure 10(a) of Riley, representing the experimental approximation proportional to the probability density function. The small circles represent the data bins. The left-most circle probably was not included by the author in the calculations of the "best fit" lines, and he may have omitted the right-most point for calculation of the "least squares" dashed line. (See text and footnote 14 for more explanation.) The dashed line in Fig. 1(a)(i) is Riley's claimed least squares (LS) line of best fit to the data. Subfigure 1(a)(ii) replaces the LS line with a dotted line whose slope is Riley's presumed MLE estimate of (−3.0) (see text). Subfigure (iii) superimposes (i) and (ii) for comparison of the dashed and dotted lines.

Riley's paragraph [42] states:

> "The MLE fit to the line in Fig. 10(b) gives a slope of −2.0."

Does this mean for a pdf $p(x) = Cx^{-\alpha}$, that $\alpha = 2.0$ or $\alpha - 1 = 2$ (so $\alpha = 3.0$)? I now conjecture that it means that $\alpha = 3$. This is because the syntax here, though ambiguous, is somewhat less ambiguous than elsewhere. Also, assuming $\alpha = 3$ along with an assumption $x_{min} = 2.6(\times 10^9)$ [15] yields a final probability

with those explanations.

The paper does not say why why the rightmost point would be omitted from the least squares calculations, if it was. My independent calculations from the original data yield slightly different numbers than the the paper's. The differences would not affect our discussion. It is possible that I was using different assumptions because the paper's assumptions are not clearly stated. In particular, the lower limit x_{min} (as defined in my earlier analysis [3]) for the power law pdf is not specified.

Since all this is probably as confusing to the reader as it was to me, I emphasize that *all* the data used in constructing our Fig. 1 came directly from Riley's paper. Because the coordinates of the data points were not directly given there, they were measured from a high-resolution .pdf file of the paper obtained from the Space Weather website.

[15] This assumption was obtained from measuring a scale value of 2.6 for the leftmost dashed vertical dashed line in Riley's Fig. 10(a) which is "as Figure 2", combined with the caption of Fig. 2. That caption states that the leftmost dashed vertical line marks the lower bound for the MLE estimate, which for mathematical correctness should be x_{min}.

estimate of 2.7% (for the probability of an event worse than Carrington in a decade), which is not too far from the paper's claim of 3.0%.

Whatever the case, Table 1 gives the measured "MLE slope" of Fig. 10(b) (graph of the CCDF) as 2.0, which would imply $\alpha \approx 3.0$.

Figure 1(a)(ii) superimposes a dotted line of slope -3.0 over the data. The constant term in the equation of that line was determined by minimizing the least squares error between the data and all lines of slope -3.0.

Figure 1(a)(iii) superimposes that dotted line on Figure 1(a)(i) in order to compare how well the two lines fit the data. The difference does not appear enormous, but I would be surprised if anyone would claim that the dotted MLE line fits the data better.

This seemingly small difference translates into a difference of about an order of magnitude in the final probability estimates (of an event worse than Carrington in a decade). Using the same assumptions as above, but changing $\alpha = 3.0$ to $\alpha = 2.0$ changes the previous probability estimate of 2.7% to 17.9%.

5.1 How these probability estimates were obtained

Many readers may want to skip this subsection, which explains the subsidiary assumptions used to calculate the probability estimate of 2.7% for $\alpha = 2.0$ vs 17.9% for $\alpha = 3.0$. Its interest lies in clarifying how to guess the critical parameter x_{min} which the paper fails to give for all the data sets.

The values substituted into the paper's equation (6) to obtain these estimates are are:

$$x_{min} = 2.58, \quad N = 55, \quad \tau = 383, \quad \Delta t = 10, \quad x_{crit} = 18.8 \quad .$$

According to McCracken, et al. [2] which presented the original data used by Riley and here, the time span τ of the data set is 383 years, not the 450 years which Riley states in paragraph 40. The number of observations at least x_{min} is only 55 out of the total of 70 observations reported by [2], despite the impression given by the paper that all 70 observations at least $2.0(\times 10^9)$ were used.[16] The paper does not explicitly give the value of x_{min}, which appears only implicitly in the expression "$P(x \geq x_{crit})$" of the paper's equation (6). The above value $x_{min} = 2.58$ was deduced from the captions of the paper's Figure 10, which refers back ("As Figure 2") to the caption of Figure 2. The latter caption states:

[16] The paper does say that all 70 observations were used for Fig. 10(b), but this is mathematically incorrect unless $x_{min} = 2(\times 10^9)$, which contradicts apparent assumptions (particularly, the captions for Figs. 10 and 2) stated elsewhere. For reference, were $x_{min} = 2$, the final probability estimate for $\alpha = 3$ would be 2.0% instead of the 2.7% for the $x_{min} = 2.58$ which is the main text's best guess.

The difference for these particular estimates is small, but they do illustrate the difficulty in understanding Riley's assumptions as presented in the paper. All of this stems from the paper's incorrect equation (1) which omits the critical parameter x_{min}, at which the reader has to guess for all of the data sets. This is one more illustration why Riley's probability estimates should be carefully checked before citation.

> "The solid straight line in Figure 2b is a MLE fit to the data above the lower threshold indicated by the left-most vertical dashed line."

For mathematical correctness, this would imply that x_{min} would be the horizontal scale value of the left-most vertical dashed line. That scale value was measured as 2.58 for Fig. 10(a) on the high-resolution .pdf file.

6 The Solar Flare data set–Part 2

The preliminary analysis of the Solar Flares section of Riley's paper focused on the inconsistency between the slopes of Figs. 2(a) and (b), which is the main mystery affecting all of the data sets. The present section reports an unrelated major anomaly. Although it does not affect the final probability estimates, it is too severe to be ignored.

The solar flare measurements report the flux of "hard" (i.e., energetic) X-rays. We can imagine observing a plate of some given size (say 2000 cm^2) and counting the number of hard X-ray photons passing through it in some given period (say a second). Observe for a longer period (say a day). Then plot the photon counts per second vs. the number of times in a day that this count frequency is observed. For example, if there are 20 periods of 1 second in the day for which the count frequency is in a "bin" of determined size, say $[10^4, 10^4 + 10]$, this would be represented on a histogram as a vertical bar over the interval $[10^4, 10^4 + 10]$ with height 20. (Riley's histograms represent the vertical bars with small circles.) To obtain Figure 2(a), plot the histogram in log-log coordinates. The result is an experimental approximation proportional to the probability density $p(x)$ of observing count frequency x, plotted in log-log coordinates. This explanation is a bit oversimplified, but it gives the physical picture.

I chose the bin width as 10 for illustration, but this is far from what Riley used. Were the bin width 10, there would something like 10^7 bins, a huge number for a data set comprising only 7236 observations [paragraph 22]. According to the caption for Figure 2(a),

> "100 bins were equally spaced in peak rate between 10^2 and 10^8 counts¯s [sic] per 2000 cm^2".

(I wonder if the 10^8 could be a typo which should be 10^6, but even if so, the following remarks would apply.) That would imply that each bin was of the enormous size $(10^8 - 10^2)/100 \approx 10^6$. In particular, the horizontal interval $[10^2, 10^4)$ would lie entirely inside one bin. But there are a large number of data points (small circles) above this interval, and each data point is supposed to correspond to one bin.

This is a huge anomaly. Something is seriously wrong. My first thought was that the 100 bins might have been equally spaced in "logarithmic" rather than "linear" space, but paragraph [23] explictly rules out this possibility:

> "These data were binned in intervals separated equally in linear, not logarithmic space . . . ."

Could this anomaly have something to do with Riley's incorrect claim (discussed in Solar Flares–Part 1) concerning the slope of Figure 2(a)? Could logarithmic bins have been used even though the paper specifically asserts the opposite?[17] I do not know.

7 The Geomagnetic Storm data set

The most important thing to recognize concerning Riley's analysis of the Geomagnetic Storm data set is that it is *impossible to replicate.* We have to take the author's word for the correctness of the analysis.[18] We cannot check it for ourselves. Given the very substantial and demonstrable errors elsewhere in the paper (particularly in the Solar Flare section), it seems prudent to withhold judgment on an analysis which cannot be independently checked.

The reason that the paper's analysis is impossible to check is that it never gives us the precise definition it uses to identify a "geomagnetic storm". The closest the paper comes is in paragraph [37]:

> "To generate an event-based data set, we arbitrarily define a 'significant' magnetic storm to be one for which $|Dst|$ exceeds 100 nT. In principle, we could define an 'event' as the hourly value of Dst and compute our probabilities based on that.
>
> ...
>
> Thus, we would rather identify a contiguous range of data that all exceeds some criteria as a single event, rather than a set of events. In Figure 7, we show the occurrence of these significant storms as a function of time. ... "

But it never specifies the "some criteria" which it uses to define "significant" single event! A reference to the raw data is furnished in the caption to its Figure 6, but there is no way to tell how the paper translates this raw data into a sequence of "significant" magnetic storms.

That said, let us proceed to discuss the paper's analysis under the assumption that its graphs, etc., can be regarded as accurate (unlike those of its Solar Flare section). Table 1 shows that this data set is different from the others in that the slope of the observed pdf (Figures (a)) is not about the same as the slope of the CCDF (Figures (b)).

Under the paper's power law assumption, the absolute "slope" α of the pdf should be precisely one more than the absolute "slope" $\alpha - 1$ of the CCDF. This

[17] If logarithmic bins were used to construct Fig. (a), but Riley's "counting to the right" method used for Fig. (b), then one can show that assuming a sufficiently large sample from a perfect power law, Figs. (a) and (b) would have the same slope. However, Fig. (a) would no longer approximate the log-log graph of the pdf.

[18] This is not meant to imply that the author is untrustworthy. But anyone can make a mistake. Remember that we are talking about probability estimates which if taken seriously, have societal implications in the hundreds of billions of dollars. Mistakes which might seem trivial in other contexts can have huge consequences in this one.

is not the case for the geomagnetic storms—the absolute slope of Figure 8(a) is about 0.7 more than the absolute slope of 8(b)—, but it seemed conceivable that the discrepancy (0.7 vs the expected 1.0) might be attributable to experimental imperfections or random chance.

To test this possibility, I wrote a computer program which simulated the paper's geomagnetic storm data, assuming it was randomly drawn from a power law distribution $p(x) = Cx^{-\alpha}$ with $\alpha = 4.2$. That is the value of α which seems to be implied by the paper's ambiguous paragraph [38] claim,

> "In Figure 8a, we show a histogram of events as a function of the severity of the storm. The data appear to follow a power law distribution, as indicated by a least squares fit to the data.[19] In Figure 8b the power law relationship of the CCDF is considerably better: Only the last 3 points (which are made up of only 1, 2, and 3 events, respectively, deviate. *The slope of the MLE fit is* -3.2. [emphasis mine]"

This is a good example of the paper's ambiguous syntax. Does "MLE fit" refer to the $-\alpha$ of the pdf, or $-(\alpha - 1)$ of the CCDF?[20]

Both Figures 8(a) and 8(b) are mentioned in this quote, but Figure 8(b) is mentioned last. Should we take this to mean that $\alpha - 1$ is the slope of Figure 8(b), where α is given by the MLE estimate? My current best guess is that this probably is the intended meaning, particularly since the measured slope -3.0 of the MLE line in Figure 8(b) is not too far from the claimed "MLE fit" of -3.2 of the quote (though still outside the maximum estimated measurement error).

Before describing the computer results, let us examine some consequences of the hypothesis that the MLE estimate for α is $3.2 + 1 = 4.2$, along with other reasonable hypotheses about the paper's ambigous presentation. In this analysis, please keep in mind that there is no way to independently reproduce the paper's assertions from the raw data, as noted above.

First let us calculate the probability of an event worse than Carrington in a decade. For this calculation, we need the values of the parameters to substitute in the paper's equation (6):

$$P(x \geq x_{crit}, t = \Delta t) = 1 - e^{-N\frac{\Delta t}{\tau}P(x \geq x_{crit})}. \qquad \text{Riley's (6)}$$

Some of these are given explicitly in the paper:

$$N = 746, \quad \tau = 46, \quad x_{crit} = 850, \quad \Delta t = 10 \quad ,$$

[19] It is too scattered to look that way to me.

[20] Before learning how to accurately measure slopes on a high-resolution .pdf file under high magnification, my best guess was $\alpha = 3.2$. My current best guess is $\alpha - 1 = 3.2$ because the measured value of $\alpha - 1 = 3.0 \pm .05$ for the slope of Fig. 8(b) (see Table 1) is closer to 3.2 than is the measured value $\alpha = 3.7 \pm .06$ for the slope of Fig. 8(a).

The arithmetic of the previous "Analysis" [3] was based on the previous best guess of $\alpha = 3.2$. That best guess was based on an indication that similar ambiguous syntax in the CME section of the paper should be interpreted in that way. A direct question of which was meant for the CME data was ignored by the author, as have been all of my four messages asking about various points in [1].

but to compute

$$P(x \geq x_{crit}) = \left(\frac{x_{min}}{x_{crit}} \right)^{\alpha-1} \quad ,$$

we need the value of x_{min}, which the paper never gives explicitly for any data set.

From the caption of Figure 2 (figures for the other data sets refer back to Figure 2 via the phrase "as Figure 2"), we can guess that x_{min} is the horizontal scale value of the "left-most vertical dashed line" in Figure (b). This caption states that

> "The solid straight line in Figure 2b is a MLE fit to the data above the lower threshold indicated by the left-most vertical dashed line."

Then for mathematical correctness, x_{min} should be the scale coordinate of the "left-most vertical dashed line".

There is also a left-most vertical dashed line in all Figures (a), but it is all but invisible in Figure 8(a) (see below). For Figure 2, this left-most line has the same scale value as that of the corresponding line in Figure 2(b), but Figure 8 is different. The scale value of the left-most vertical dashed lne in Figure 8(b) is seen in a magnified high-resolution .pdf file to be about 121 (more exactly, 121.2±1.5). In the printed version, Figure 8(a) appears to have *only one* vertical dashed line (unlike all the other figures), but under magnification of the .pdf file, the left-most vertical line is seen to be *coincident* with the left vertical axis at scale value 100. The dashes are a little wider than the vertical axis under 800% magnification.

Thus we have two candidates for x_{min}, 121 and 100. If we take the captions for Figs. 8 and 2 absolutely literally, we would have $x_{min} = 121$, but we shall do the calculation separately for each value.

Taking $x_{min} = 121$, the final probability estimate (of an event worse than Carrington in a decade, as determined from Riley's (6)) is 27.4%, which seems too high to be believable.[21] The paper claims 12%. Taking $x_{min} = 100$, yields a final probability estimate of 16%, which differs significantly from the paper's claimed 12%, but at least is in the same ballpark.

The larger α, the smaller the final probability estimate. Recall that the measured slope of the MLE line in Fig. 8(b) is -3.0, which would imply $\alpha = 4.0$. Taking into account the possible measurement error (estimated as ±.06), suppose we increase this to $\alpha = 4.1$, again using the smaller value of $x_{min} = 100$ (which will produce a smaller probability estimate). Then the final probability would be 19%, which still seems suspiciously high.

I haven't found reasonable hypotheses which yield the paper's claim of 12%. It appears that either my arithmetic or the paper's must be wrong. If any reader has an explanation, I surely would like to hear it. Our smallest estimate of 16% above is in the same ballpark as the paper's 12%, but differs too much to be attributable to any likely causes, such as measurement error.

[21] Were that the case, the probability of going more than 15 decades (from 1859 to 2015) without an event as bad would be less than $(1 - .274)^{15} < .008$.

This, along with the serious, demonstrable errors elsewhere in the paper and the impossibility of checking Riley's claims from the raw data (because we lack his definition of "significant" magnetic storm) should be taken into account by anyone tempted to cite the paper without caveats. I think it would be irresponsible to cite the paper's final probability estimates without caveats unless the citor has independently verified and is willing to defend them.

7.1 Computer experiments

The computer experiments to be described were a product of a measurement oversight. (Yes, I make mistakes, too! But I do try to correct them when discovered.) In all of the paper's Figures *except* Figure 8, the vertical scale consists of an integral number of powers of ten (e.g., runs from 10^0 to 10^3 for Fig. 4). However, Fig. 8(b) runs from about $.003 \approx 10^{-2.5}$ to $1.00 = 10^0$ (instead of the expected 3 orders of magnitude from $.001 = 10^{-3}$ to $1.00 = 10^0$). Not noticing this and thinking that the vertical scale was three orders of magnitude, I obtained a slope of Fig. 8(b) close to that of Fig. 8(a), just like the other three figures in which the slopes of Figs. (a) were about the same as those of the respective Fig. (b).

It seemed remarkable that *all* the Figs. (a) had about the same slopes as the corresponding Fig. (b) when they should differ by 1, given that they were constructed independently from data sets describing completely different physical phenomena. (Indeed, the ice core data set probably has nothing to do with anything involving the sun!)

I wondered if some property of power law distributions might make least squares (LS) estimates of slopes of Figs. (a) systematically one less than MLE estimates of the same slope. Indeed, such a phenomenon was reported by Goldstein, et al. [6] (their Table 1), although under significantly different protocols than used by Riley. Specifically, they constructed the LS estimates from *all* observations drawn from a pure power law, instead of restricting to observations in a limited range (between the dashed vertical lines of Riley's Figs. (a)) as Riley does. There are reasons to expect that using all observations might decrease the LS estimates of the observed slope compared to restricting to a range of observations for which nearly all observation "bins" were nonempty.

So, I wrote a computer program to simulate Riley's "experiment" of drawing 746 observations from a pure power law distribution $p(x) = Cx^{-\alpha}$, with $\alpha = 4.2$ corresponding to my current best guess at Riley's claim. The object was to see if the LS estimates for α were systematically less than the MLE estimates.

This computer experiment analyzes Fig. 8(a), so it used the parameters of that figure. Specifically, the LS estimate was calculated by using only observations between the left-most vertical dashed line (visible only under high magnification at scale value 100 as described above) and right-most vertical dashed line (at scale value 200). For consistency, x_{min} must then be taken as 100.

The caption of Figure 8 states:

> “In Figure 8(a), 100 bins were equally spaced in Dst between $|Dst| = 100$ and $|Dst| = 10^{2.7}[\simeq 501.2]$.”

This implies that there were

$$100 \times \frac{200 - 100}{10^{2.7} - 100} \simeq 25$$

bins beween 100 and 200, the range which is used to calculate the least squares estimate LS for α (the negative of the slope of Fig. 8(a)). For some samples, some of these bins will be empty and therefore not shown on Fig. 8(a). (Because that figure is in log-log coordinates, an empty bin would be represented by a small circle at vertical coordinate $-\infty$.) Thus there should be no more than 25 circles on Fig. 8(a) between horizontal coordinates 100 and 200. However, a direct count from the .pdf file at high resolution shows 32 filled bins (i.e., 32 bins with nonempty contents). This anomaly is unexplained.

It seemed artificial to use the 25 bins between 100 and 200 claimed by the caption of Fig. 2(a) when the data from which the LS estimate for α was calculated used at least 32 bins. So, the program used 32 bins of equal width between 100 and 200, under the assumption that the bounds 100 and 200 were were chosen so that nearly all bins would be filled. (In the computer runs, this was the case—it was rare that even one of the 32 bins would be empty.)

The program drew 200 samples of 746 observations each from this pure power law and computed both LS and MLE estimates for α.[22] The difference between the measured slopes of Riley’s Figures. 8(a) and 8(b) is 0.7, with 1.0 expected for a pure power law. This corresponds to the difference MLE − LS for Fig. 8(a) being larger than 0.3.[23] The program calculated the MLE and LS estimates for each of the 200 samples of size 746, and flagged those with MLE − LS > 0.3.

Of the 200 samples only 6 had MLE − LS > 0.3. This corresponds to a probability of about 3% that a sample of 746 drawn from Riley’s (apparently) assumed power law will result in the MLE estimate being as much as 0.3 higher than the LS estimate.

The use of the one-tail test seems reasonable because for all of the data sets’ Fig. (a), MLE > LS, and for all but the geomagnetic storm data, MLE is substantially higher (about 1 higher) than LS. The question is: why do we always observe MLE > LS when they should be nearly equal? Could this be attributed to some statistical feature of power laws?

[22] The MLE estimates, of course, used all observations, not just those below horizontal scale value 200. It would be mathematically incorrect to restrict to scale values no more than 200. Also, the coincidence between the number of samples of size 746, which was 200, and the scale value 200 is no more than a coincidence. The program took 20 samples in each run, and there were 10 runs. After the data were collected and the coincidence noticed, it seemed artificial to restrict, say, to just 5 runs in order to remove the coincidence.

[23] This is a bit confusing because Riley’s paragraph [38] claim of “MLE fit” of -3.2 for (presumably) Fig. 8(b) is inconsistent with the measured slope of -3.0 for the solid MLE line of Fig. 8(b). If Riley had claimed the measured -3.0, that would correspond to $\alpha = 4.0$. but the LS line of Fig. 8(a) has absolute slope of only 3.7 (see Table 1). So for Fig. 8(a), MLE − LS = 4.0 − 3.7 = 0.3. But $\alpha = 4.2$ was used in the power law because that seems to be Riley’s claim.

I had expected that MLE $-$ LS > 0 would be the norm, partly because that was emphatically the case for the data sets (of which three of four had MLE $-$ LS ≥ 1 and partly because [6] reported MLE $-$ LS $\simeq 0.9$ (though using a substantially different protocol). But the results showed that MLE $-$ LS < -0.3 was actually much more common than MLE $-$ LS > 0.3. Of the 200 samples, 21 had MLE $-$ LS < -0.3, with only 6 showing MLE $-$ LS > 0.3.

In conclusion, if Riley's geomagnetic storm data were in fact drawn from the pure power law $p(x) = Cx^{-4.2}$ that he seems to assume, his particular sample was rather anomalous. More specifically, we can reject at the 97% confidence level level using a one-tail test that Riley's geomagnetic storm data was drawn from the power law distribution which he seems to assume. For a two-tailed test, we can only reject at the confidence level of 86%, which is not terribly high, but still indicates that that Riley's sample looks anomalous.

This is one reason to doubt that Riley's sample came from the assumed power law $p(x) = Cx^{-4.2}$. If it did come from that power law it was a fairly anomalous sample. Of course, if a sample is known to be anomalous, one should hesitate to draw statistical conclusions from it, such as a conclusion about the probability of an event worse than Carrington in a decade.

8 The Coronal Mass Ejection data set

The Coronal Mass Ejection data set was extensively discussed in the previous "Analysis" [3], and will not be repeated in detail here. The only change I would make is that my current best guess as to the meaning of ambiguous language concerning "MLE fit" is that it refers to the slope of the CCDF, rather than the slope of the pdf as I had previously assumed.[24] I had interpreted the paper's highly ambiguous language in paragraph [31] as meaning that in the assumed pdf $p(x) = Cx^{-\alpha}$, $\alpha = 3.2$. In the light of subsequent development's my best guess is that the paper meant $\alpha = 4.2$. That changes the final probability estimates.

Any analysis of [3] depending on $\alpha = 3.2$ should be reevaluated in this light. This includes the analysis suggesting that the paper's initial probability estimate of 85% (for the probability of an event worse than Carrington in the next decade) may have been obtained by using an incorrect $x_{min} = 100$ instead of the correct $x_{min} = 700$. I was unable to repeat the analysis using Riley's so-called "quadratic speeds", but I was able to repeat the calculation using "linear speeds" instead,[25] which should yield a comparable result. The result

[24] Before posting the previous "Analysis" [3], I wrote the author asking about the meaning. He chose to ignore that along with three other messages. If my previous best guess turned out to be wrong (which is *still* uncertain), it was not for lack of inquiry. I would advise anyone tempted to cite Riley's final probability estimates without caveats to inquire of the author as to his meaning, and if he replies, then to repeat the arithmetic.

[25] "Quadratic speeds" are obtained by fitting a quadratic polynomial to the height vs. time measurements of CME's and using the slope of the quadratic at the highest measurements. "Linear speeds" are similarly obtained, but using a linear polynomial instead of quadratic.

The reason that I could not use quadratic speeds is that the database could be searched

I obtained which is reported in [3] using $\alpha = 3.2$ was probability 99.99%. The same calculation using $\alpha = 4.2$ gives probability 88%, which is comparable to Riley's 85%.

The paper recognizes its 85% estimate as "not credible" [paragraph 31]. Let us think of the implications of this.

Both Figures 4(a) and 4(b) look like straight lines between speeds 700 and 2000 km/s. This is the essentially the *only* justification that the paper gives for the power law assumption. But if that assumption leads to a result which is 'recognized as "not credible", then it must be wrong. Yet the paper soldiers on under the same assumption to eventually obtain its famous 12% final estimate for the probability of an event worse than Carrington in the next decade.

The paper explains this as follows. Figure 4(b) looks like a straight line from 700 to 2000 km/s, but above 2000 it deviates noticeably, starting with what paper calls a "knee" at 2000:

> "We suggest that the reason for this artificially high probability is that above 2000 km s^{-1} there appears to be a well-defined "knee" in the distribution. Therefore, to address this, in Figure 5, we have replotted the CCDF for the highest speeds and computed the MLE fit to only the distribution above 2000 km s^{-1}."

When the paper acknowledges that "above 2000 [the CCDF] deviates noticeably", it effectively acknowledges that a power law assumption based only on the appearance of the pdf or CCDF as approximately a straight line is not in accord with observation. Yet it goes on to apply the power law assumption (with a revised slope) as if it were valid! Moreover, it does not tell us that the data justifying the revised assumption (i.e., speeds above 2000) consists of only a few tens of observations.

This alone should be enough to disqualify the final 12% estimate as anything other than the result of an academic inquiry. It is hard to believe that it could be taken seriously as a basis for determining, say, reasonable spending to prepare for an event that could cost trillions of dollars. Twelve percent of a trillion is $120 billion, which could be justified were the 12% estimate soundly based. Yet it apparently has been so taken, judging by comments in the popular press. Even NASA has posted an article on its website which takes seriously Riley's 12%.[26]

Many of the questionable aspects of Riley's analyses would be unlikely to be noted by casual readers. Some are too technical, and some require unusually careful reading. But this one should stand out to any qualified reader.

automatically by linear speeds, but not by quadratic speeds. To sort the data by quadratic speeds, I would have had to sort thousands of observations by hand.

[26] http://science.nasa.gov/science-news/science-at-nasa/2014/23jul_superstorm/ The article actually cites the same 12% estimate derived from the Geomagnetic Storm data, but there are serious problems with that one as well.

9 Summary and conclusions

Riley considers four data sets, Solar Flares, Coronal Mass Ejections, Geomagnetic Storms, and Nitrates in Ice Core Samples. Only the last three of these yield estimates for the probability of an event worse than Carrington in the next decade.

The raw data for the ice core samples has been questioned in Wolff, et al. [7], (of which Riley may have been unaware because it was published after Riley's paper [1]). Essentially Wolff, et al., assert that the ice core samples from which Riley's data was taken were probably misdated and also probably caused by forest fires rather than emissions from the sun.

The paper's analysis of the solar flare data appears to be seriously in error. This data was not used to produce a final probability estimate, but the errors are so serious that they make one wonder about the handling of the other data.

The paper's analysis of the Coronal Mass Ejection data does not support the conclusion of its famous estimate of 12% probability for an event worse than Carrington in the next decade. An initial estimate based on over 1000 observations of equal reliability (or unreliability) yielded an estimate of 85%, which the paper itself characterizes as "not credible" Then the paper applies identical methods to a restricted data set of only a few tens of observations to obtain the 12% estimate. If the underlying power law assumption can be trusted, then the original analysis should not have produced the unbelievable 85% estimate.

The paper's analysis of the Geomagnetic Storm data is described so vaguely that it would be *impossible* to replicate. Serious errors elsewhere in the paper's data analyses (e.g., for Solar Flares) suggest that any analyses which cannot be replicated should be regarded as provisional. Moreover, using only results reported by the paper itself together with the paper's methods, my final final probabiity estimates differ substantially from those reported in the paper. Either my arithmetic is wrong, or the paper's is.

The above points out anomalies in the individual data sets. However, there is one major anomaly which unites all of them—the "slopes" of the reported pdf's (Figures (a)) (i.e., the slopes of their log-log graphs) are inconsistent with the slopes of the CCDF's (Figures (b)), assuming the power laws on which the paper's analysis is based. For three of the four data sets, the slopes of the pdf's are about the same as the slopes of the CCDF's, when according to the power law, they should differ by 1.

For the remaining data set, Geomagnetic Storms, the difference is not nearly zero, but also is not as close to 1 as it should be.. Computer experiments suggest that a power law hypothesis for this data set can be rejected at the 97% confidence level. That is, there is only about a 3% probability that the reported results would be obtained from the paper's assumed power law.

This suggests that either the power law assumption is wrong, or the data set *was* drawn from the assumed power law distribution, but happened to be anomalous. An analogy is that if a fair coin is tossed 5 times it *could* come up all heads even thouugh the probability of this is only aboout 3%. But if the data set *is* recognized as anomalous, it is questionable to base conclusions on it.

The main conclusion that I draw from all of this is that Riley's paper, though an interesting academic exercise, should not be cited without caveats as producing realistic probability estimates. Its methods are questionable, and their implementations are often flawed. The paper should be cited without caveats only by those who have independently checked its assertions and are willing to defend them.

10 The broader picture

I am sometimes asked how I would estimate the probability of an event as bad as Carrington in the next decade. My reply is that I don't know any way to do so in which I would put any confidence.

I doubt that we have enough information. Essentially, all we know is that there has been no event like that for about 15 decades (from 1859 to 2015). If there were some reliable way to assign a probability distribution to the strength of events on the sun (e.g., Coronal Mass Ejections) or their manifestations on the Earth (e.g., geomagnetic storms as measured in Dst), Riley's method could be used (whether or not that distribution were a power law), but I've seen no indication that we are close to identifying such a distribution. In particular, there seems to be more negative evidence than positive for a power law distribution.

The fancier the mathematics leading to such an estimate, the less I would trust the estimate. Nontrivial mathematics generally requires nontrivial assumptions which are difficult to verify in a real world setting. So, let's see how far we can get with transparent mathematics using plausible assumptions.

We shall use a decade as a time unit. Let q denote the probability of an event as bad as Carrington in a given decade. This formulation already contains an implicit assumption that q does not depend on the decade—technically known as an assumption of "stationarity" (of a certain stochastic process which we're not going to define).

We shall also assume that the future is "independent" of the past. In particular, if a magnetic storm of some strength (not necessarily as bad as Carrington) happens to have occurred yesterday, that does not affect in any way the probability of anything tomorrow. Put another way, the probability of some event tomorrow is the same whether or not we know that a magnetic storm occurred yesterday.

Under these assumptions, the probability that 15 decades pass with no event as bad as Carrington is $(1-q)^{15}$. (I assume that the reader knows enough elementary probability to figure out why.) If someone claims, say, that $q = 0.12$ (i.e., 12%), then we can calculate that probability as $(1-0.12)^{15} \simeq 0.15$. If that probability is very low, then it is very unlikely that the observed 15 decades without a Carrington would have occurred, so perhaps the claimed q was too high. In technical terms, we can reject the hypothesis that $q \geq 0.12$ at the 85% confidence level. $(1-0.15 = 0.85.)$ If the claimed q were 0.18 (i.e., 18%), so that $(1-q)^{15} \simeq .05$, then we could reject the hypothesis that $q \geq 0.18$ at the

95% confidence level, etc.

A commonly used confidence level required to reject a hypothesis is 95%, but this is an arbitrary choice, and many feel that a more stringent level such as 99% or even 99.9 % should be used for important matters. But even if we can't reject a hypothesis like $q \geq 0.12$ at our chosen level, that doesn't mean that we should *accept* the hypothesis.[27] It just means that the hypothesis is not terribly unlikely, though it might not be particularly likely either.

In such a case, we would be less likely to draw a conclusion either way. In particular, Riley's famous 12% estimate seems fairly unlikely, but not terribly unlikely.

But more can be said. If we are assuming that Carrington-class events occur at random times, then we would not estimate the expected time between events as 15 decades but *at least* twice that, 30 decades. That is because our only information is that none has been seen for 15 decades, so why would we not expect *at least* another 15 decades without one (on average)?

If we take this point of view, then our calculation $(1-q)^{15}$ should be replaced by $(1-q)^{30}$ to determine the hypothesis rejection at the chosen confidence level. And even that tilts the scale *against* rejection, since we guess (estimate) the average time between Carringtons as *at least* 30 decades, instead of exactly 30 decades. So our rejection criterion is actually somewhat conservative, in that we might fail to reject some hypotheses which actually should be rejected.

Applied to Riley's 12% estimate, we get $(1 - q)^{30} = 0.88^{30} \simeq .02$, so we can reject that $q \geq 0.12$ at the 98% confidence level. Even disregarding what I view as questionable methods leading to the 12% estimate, I would distrust it on the general grounds just explained. There is no way to prove that it can't be correct, but it does seem suspect.

We have been using Riley's 12% estimate for illustration, but even under the most optimistic expectations, the uncertainties in arriving at such estimates are so large that it would be unrealistic to hope for accuracy much better than an order of magnitude. So let's repeat the calculation for a q which is an order of magnitude lower, say $q := .01$ (i.e., 1%). Then $(1 - q)^{30} = 0.99^{30} = .74$, which is such a large probability that we couldn't think of rejecting the hypothesis $q \geq 0.01$ at any reasonable confidence level.

The bottom line is that it seems fairly unlikely that q could be as high as 0.1 (10%) because $(1 - 0.1)^{30} \simeq .04$. But if q were .01, the observed 15 decades without a Carrington (followed by a similar presumed 15 decades without one, on average) would be very probable, so we certainly can't reject $q \geq .01$. That doesn't mean that we have to *accept* or estimate that $q \geq .01$. It could well be much lower, but erring on the side of caution seems prudent to me, given the potentially catastrophic consequences of a Carrington class event.

Such an event could disrupt life as we know it for a year or more and could cost in the trillions of dollars. A percent of a trillion is 10 billion, so we would more or less break even if we could substantially reduce the threat by spending on the order of tens of billions. That amounts to some tens of dollars per

[27] Many elementary statistics texts give the opposite impression.

inhabitant of the U. S. . That sounds to me like a reasonable amount for insurance, *assuming* that good protection could be obtained for that amount.

11 Appendix 1: The Maximum Likelihood Estimate (MLE)

11.1 Synopsis

The Maximum Likelihood Estimate (MLE) plays a large role in Riley's analysis. Any misunderstanding of it may engender a misunderstanding of the paper. This appendix reviews this concept for those not familiar with it. The points made will be the following:

- Given a family of probability density functions (pdf) depending on a parameter α, such as an exponential distribution $p_\alpha(x) = \alpha e^{-\alpha x}$, and a sample from some distribution whose pdf is known to be in the family (but for an unknown α), the MLE gives an estimate of α.

  When one refers to an MLE in standard statistical terminology, one refers to the estimate of the parameter for a family of pdf's. There is no usual concept of an "MLE" for a family of complementary cumulative distribution functions (CCDF). Given the pdf's depending on α and the sample, there is a way (usually yielding a formula) to determine the MLE of α, but there is usually no straightforward way to estimate α given only a family of CCDF's. Riley appears to use nonstandard terminology in talking about the "MLE fit" [paragraph 38] and the like for a family of CCDF's depending on a parameter.

  Riley's apparent terminology could be justified for the special case of *power law* CCDF's if explicitly introduced, but it never is. This makes the syntax ambiguous and has the potential to cause great confusion.

- Riley's formula (4) for the MLE applies *only* to power law pdf's. The paper does not say otherwise, but a reader who did not keep this in mind. might be tempted to draw inappropriate conclusions.

  If the formula were applied to a family of pdf's other than power laws, one would not expect sensible results. This is important in the context of the present analysis, which presents evidence that Riley's data may *not* have come from power laws. The MLE's calculated from Riley's (4) do not always appear to well fit the observed pdf's, and this may be the reason.

11.2 Review of the Maximum Likelihood Estimate (MLE)

Suppose we have a family of probability distributions indexed by a real parameter α. We take a random sample from the distribution and try to use it to estimate the parameter.

For a simple example, suppose we observe an asymetrical coin tossed five times, with a result of three "heads", and we are asked to estimate the probability α that another toss will give a "head". In the absence of additional information, a natural guess is that α might be the value which maximizes the probability of obtaining the observed result. The latter probability is proportional to $\alpha^3(1-\alpha)^2$, which is maximum when $\alpha = 3/5$.

The $3/5$ is called the Maximum Likelihood Estimate (MLE) for α. For a discrete probability space, it is defined as the value of the parameter for which the observed sample is most probable. For a real number x drawn from some interval of real numbers with probability density function $p_\alpha = p_\alpha(x)$, the probability of any finite sample $x_1, x_2, \ldots, x_n$ of n numbers drawn independently is zero, so we replace that probability with the "likelihood" $L(x_1, x_2, \ldots, x_n; \alpha)$ of the sample, defined by:

$$L(x_1, x_2, \ldots, x_n; \alpha) := p_\alpha(x_1) p_\alpha(x_2) \ldots p_\alpha(x_n). \tag{1}$$

For a given sample $x_1, \ldots, x_n$, the value of $\alpha = \alpha_{MLE}$ which maximizes $L(x_1, \ldots, x_n; \alpha)$ is called the *maximum likelihood estimate* (MLE) for α.

For a power law distribution

$$p(x) = \begin{cases} C_\alpha x^{-\alpha} & x \geq x_{min} \\ \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $C_\alpha = (\alpha - 1)x_{min}^{\alpha-1}$ is a constant depending on α, the likelihood function is

$$L(x_1, \ldots, x_n; \alpha) = C_\alpha^n (x_1 x_2 \ldots x_n)^{-\alpha} \quad .$$

Maximizing this with respect to α for a given sample $x_1, \ldots, x_n$ gives the Maximum Likelihood Estimate α_{MLE}:

$$\alpha_{MLE} = 1 + \left[\sum_{i=1}^{n} \log_e(x_i / x_{min}) \right]^{-1} \quad ,$$

which is Riley's equation (4) slightly rewritten.

All of this is of course elementary, and I summarize it only to make the point that Riley's (4) is not the fundamental definition of "maximum likelihood estimate", but applies only to the special case of a power law distribution. If it is applied to data drawn from a parametrized distribution which is *not* a power law, then one would not expect the result to have anything to do with the MLE for that distribution. For example, consider a normal distribution with mean α and variance 1, whose pdf p(x) is

$$p(x) := \frac{1}{\sqrt{2\pi}} e^{-(x-\alpha)^2/2} \quad .$$

For this distribution, maximizing the likelihood gives the very different formula

$$\alpha_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad .$$

Here if the sample values x_i tend to be large, then the MLE estimate for α also tends to be large, whereas for Riley's (4) for a power law, it is just the opposite.

My suspicion is that the data did *not* come from a power law. If it did come from a power law, we would expect the slope of Riley's Figures (b) to be one less than the slope of the corresponding Figures (a), but in fact they are nearly equal for three of the four data sets. We cannot rule out that this could happen for an atypical sample, just as we cannot rule out that a fair coin tossed 100 times could produce 100 heads. However, it seems implausible that this should happen "by accident" for three of four data samples which were not only independently taken by others, but involved different physical phenomena (X-rays from solar flares, coronal mass ejections, magnetic storms, and ice core samples). And if it did happen, any conclusions based on the samples should be suspect, precisely because the samples were known to be atypical.

Next we use the example of the variance-1 normal distribution to make another point. The complementary cumulative distribution function (CCDF) F for this normal distribution is

$$F(x) := \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-(u-\alpha)^2/2}\, du$$

It is known that there is no formula for F in terms of "elementary" functions. Like the pdf from which it was derived, it contains a parameter α. What should one make of a statement like:

> The MLE fit to F is -3.2.?

Riley contains several statements with similar syntax (e.g., pararaphs [38] and [42]). The above is not a direct quote but is representative; the actual quotes are embedded in other syntax, which makes it awkward to give a concise actual quote.

The only sensible interpretation I can think of is that the MLE estimate for α, derived from the pdf $p(x) := e^{-x^2/2}/\sqrt{2\pi}$ and sample $x_1, \ldots, x_n$, is $\alpha_{MLE} = -3.2$. There is no recognized concept of "MLE estimate" or "MLE fit" to a CCDF F. There is no obvious way to obtain an estimate for α directly from F other than to to estimate p from F and then obtain the MLE estimate from maximizing the likelihood (1). However, it now looks as if Riley may have used a different meaning.[28]

The situation becomes potentially ambiguous when applied to a power law distribution (2) with exponent α. The CCDF F for this distribution just happens to follow a power law

$$F(x) := \int_x^\infty C_\alpha u^{-\alpha}\, du = \left(\frac{x}{x_{min}}\right)^{\alpha-1} \quad ,$$

with exponent $\alpha-1$ one less than the exponent α of the pdf $p(x) = C_\alpha x^{-\alpha}$. Here it may be ambiguous whether an "MLE estimate for F" refers to an estimate

[28] A direct query to the author as to the meaning went unanswered, as have all my questions about various points in the paper.

for α or $\alpha - 1$. The MLE method estimates α, but of course that also estimates $\alpha - 1$. Riley uses vague syntax which is often ambiguous as to whether α or $\alpha - 1$ is meant.

12 Appendix 2: Measuring the .pdf file

Each data set contains references to "MLE fits" undefined "slopes" and the like. Most of these references are phrased in such a vague way that it's not clear if the author is referring to the "slope" α of the pdf $p(x) = Cx^{-\alpha}$ or the "slope" $\alpha - 1$ of the corresponding CCDF. Early in my study of this paper, I wrote the author asking which he meant, but this message was ignored, as were all of my four messages asking about various points in the paper.

To settle the question of whether the various references to "slopes" referred to the pdf or CCDF, I tried to actually measure with a ruler the slopes of the figures (a) (pdf) and (b) (CCDF), but the lines are thick enough that I couldn't get sufficient resolution to be sure.

Then David Roodman showed me how to use a high-resolution .pdf copy of the paper (obtainable from the Space Weather website) to measure with Adobe Acrobat. Using Adobe's magnification, it is easy to measure slope to an accuracy of 0.1 or better.

However, it required some experimentation to learn how. This appendix describes how the measurements were made. It may be useful for those who want to make their own measurements, and for those who want to convince themselves that my measurements were made correctly. No doubt explanations of the measurement facilities of Adobe Acrobat exist somewhere on the Adobe website, but I have not seen them, and they are not self-evident.

The measurement procedure is slightly tricky because of the use of log-log coordinates and different scales on the horizontal and vertical axes. The following instructions apply to Adobe Acrobat as implemented in Windows XP. I assume they would be similar for other common operating systems.

Get a high-resolution copy of Riley's paper from the Wiley website given in footnote 13. (So far as I know, any copy from that website is high resolution.) Following are some hints on measuring it.

1. Choose a magnification from the menu on the tool bar just above the text. For most purposes, 400% magnification is adequate, but to measure slopes above 5 accurately (which only occur in Riley's Figure 5), you may need to go above 1000%.

2. Enable the magnification tools from the edit menu by Edit > Analysis > Measuring tool. At high magnifications, the lines will appear too thick to measure accurately. They can be made thinner by unchecking "Line Weights" in View > Show/Hide > Rulers and Grids > Line Weights.

3. To make many numerical measurements, you need a scale, called a ruler. The scales are enabled by *right-clicking*, which brings up a menu (distinct

from the usual Edit, View, etc. menus). Then click on “Show Rulers”. That will present rulers on the top and left side of the page.

4. The default ruler is calibrated in inches with the usual $1/4, 1/8, \ldots$ inches. If you prefer different units, right click on the ruler itself to obtain them. I use millimeters (mm).

Adobe furnishes various measuring tools but the most straightforward and accurate procedure is to take measurements directly from the rulers. When you position the cursor on the page, the corresponding horizontal and vertical coordinates are displayed on the top and side rulers. The scale on the side ruler (y-axis) is unusual in that it increases in the *downward* direction rather than upward as is customary. For our purposes, this isn't an inconvenience because all slopes in Riley are negative and involve only coordinate *differences.*

Riley's Figures 2(a) and (b) represent the data in log-log coordinates. This means that a given difference in physical distance on the horizontal axis represents increasing the data by a constant factor. An increase by a factor of 10 corresponds to an increase in physical distance by a certain amount. It is natural to choose that amount as a “unit distance”. If so, physical distance on the horizontal axis is measured in what we call “horizontal distance units”, abbreviated *hdu.* Starting at a point and going to the right one hdu corresponds to increasing the number on the horizontal scale by a factor of 10.

Similar remarks apply to “vertical distance units” (abbreviated *vdu*) on the vertical scale. However, the physical distance (in mm, say) corresponding to 1 hdu is not usually the same as the distance corresponding to 1 vdu. Therefore what *appears* to be the slope of a line in Riley's figures, which I call the “apparent slope”, is usually not the same as the “actual slope” which would be the “apparent slope” if an hdu were the same number of millimeters as a vdu. The apparent slope is what is measured by Adobe's rulers and also what is perceived by the human eye as “steepness”. A conversion factor must be applied to obtain the actual slope.

To illustrate, let us measure the slope of the dashed LS line in Riley's Figure 2(a). First we obtain the ruler coordinates (in mm) of two points on the line, such as (40.0, 50.5) and (48.0, 71.0). For best accuracy, these points should be chosen as far apart as convenient. The apparent slope is then

$$\text{apparent slope} = -\frac{71.0 - 50.5}{48.0 - 40.0} = -2.43 \quad .$$

The minus in front of the fraction is to compensate for the fact that the vertical ruler increases in the downward direction, rather than upward as is conventional. We give numerical results to three significant figures to reduce round-off errors. At 400% magnification, the rulers can be read to about ±0.1 mm, which we shall see results in a maximum error in apparent slope of 0.13 and a maximum error in actual slope of about ±0.05. Errors in actual slope of up to about 0.2 have no practical significance.

To convert to actual slope, we use the log-log scale printed in the article to convert from hdu or vdu to mm. We find that

$$\begin{aligned} 6 \text{ hdu} &= 26.5 \text{ mm} \\ 3 \text{ vdu} &= 38.5 \text{ mm} \quad , \end{aligned}$$

from which it follows that

$$\begin{aligned} 1\text{(horizontal) mm} &= 6/26.5 = 0.264 \text{ hdu} \\ 1\text{(vertical) mm} &= 3/38.5 = 0.0792 \text{ vdu} \quad . \end{aligned}$$

Plugging this into the formula for apparent slope (converting the numerator of the fraction from mm to vdu and the denominator to hdu), gives

$$\begin{aligned} \text{actual slope} &= \text{apparent slope} \times \frac{3/38.5}{6/26.5} \\ \text{actual slope} &= \text{apparent slope} \times 0.344 \\ &= -0.84 \quad . \end{aligned}$$

We shall later need a name for the conversion factor $(3/38.5)/(6/26.5) = 0.344$ from apparent to actual slope . Call this γ, so

$$\text{actual slope} = \text{apparent slope} \times \gamma \tag{3}$$

A similar measurment of Figure 2(b) yields an actual slope of -0.85 for the dashed LS line and -0.84 for the solid MLE line. Thus the slopes of Figures 2(a) and 2(b) are effectively identical.

12.0.1 Error estimates

The maximum error in the slope estimate can be bounded as follows. First we need a crude guess at an upper bound B for the absolute (actual) slope to be measured. For Figure 2(a), we can take this as slope 2.0, which is above Riley's claim of 1.8 (but see below for my general method of determining B). Then an upper bound for the apparent slope is B/γ.

For simplicity of exposition, we assume that a positive slope m is being measured to avoid having to add confusing qualifiers concerning signs. We use the usual formula

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

where (x_1, y_1) and (x_2, y_2) are points on the line For simplicity we assume that $x_2 > x_1$ and $y_2 > y_1$, and that the y-coordinate increases in the upward direction as is usual.

Suppose that any distance measurement is subject to a maximum error of $\pm\delta$ with $\delta > 0$. At 400% magnification, δ is about 0.1 mm. Let

$$Y := y_2 - y_1 \quad \text{and} \quad X := x_2 - x_1 .$$

Then both X and Y are subject to maximum errors of $\pm 2\delta$. To keep everything positive for simplicity, we assume that both X and Y are at least 2δ.

The maximum possible error Δm in the measured (apparent) slope will occur when the numerator is as large as possible and the denominator is as small as possible:

$$\begin{aligned}
\Delta m &= \frac{Y+2\delta}{X-2\delta} - \frac{Y}{X} \\
&= \frac{2\delta(X+Y)}{X(X-2\delta)} \\
&= \frac{2\delta(1+Y/X)}{(X-2\delta)} \\
&\leq \frac{2\delta(1+B/\gamma)}{(X-2\delta)} \quad ,
\end{aligned}$$

where γ is the conversion factor from apparent slope to actual slope defined in equation (3).

For the Fig. 2(a) measurement above, with $B := 2, \delta = 0.1, \gamma = 0.344$, and $X = 8.0$ we have the maximum error

$$\Delta m \leq 0.17 \quad .$$

This is the maximum error in *apparent* slope. The maximum error Δm_{actual} in actual slope is

$$\Delta m_{actual} = \gamma \Delta m \leq .06$$

The upper bound $B := 2$ for the actual absolute slope was used above for expository convenience, but the error estimates quoted in Table 1 used an even more conservative B. The apparent slope was measured and converted to an actual slope. The bound B was then taken as the least integer at least 0.5 higher than the actual slope. For example, if the actual slope was -2.4, B was taken to be 3.0, and for actual slope -2.6, $B := 4.0$.

The goal was to make all maximum errors no more than ± 0.1. Typical errors are probably much less.

12.0.2 A shortcut to slope measurement

The above measurement method is straightforward but a bit tedious. Before leaving the subject of measurement, we mention a shortcut which readers may want to employ. One of the measurement tools furnished by Adobe can measure the "angle" θ of a line directly. The slope is then $\pm \tan \theta$.

I can't describe it further because the way it works is hard to describe in words. The easiest way to learn it is to ask someone to show you. It seems a bit strange at first, but is fairly straightforward once you get the hang of it. It produces acceptable slope estimates very easily, but is not quite as accurate as the direct measurement described above. All the "measured slopes" in this document were produced by direct measurement.

Acknowledgement: I thank David Roodman for introducing me to measurement of .pdf files, for drawing my attention to [7], and especially for sharing his independent analyses of Riley's data.

References

[1] P. Riley, "On the probability of occurrence of extreme space weather events", *Space Weather* **10** (2012), S02012

[2] K. G. McCracken, G. A. M. Dreschhoff, E. J. Zeller, D. F. Smart, and E. A. Shea, "Solar cosmic ray events for the period 1561-1994 1. Identification in polar ice, 1561-1950", *J. Geophys. Res.* **106** No. A.10 (2001), 21,585-21,598

[3] http://math.umb.edu/∼sp/analysis.pdf

[4] B. R. Dennis, "Solar hard X-ray bursts", *Sol. Phys.* **100** (1985), 465-490

[5] Lin, R. P., Schwartz, R A., Kane, S. R., Pelling, R. M., and Hurley, K. C., "Solar hard X-Ray microflares", *Astrophys. J.,* **283**, 421-425

[6] Goldstein, M. , Morris, S. , and Yen, G., "Problems with Fitting to the Power-Law Distribution", *Eur. Phys. J., 41* (2004), 255-258, arXiv: cond-mat/0402322

[7] Wolff, E. W., Bigler, M., Curran, M. A. J., Dibb, J. E., Frey, M. M., Legrand, M. , and McConnell, J. R. , "The Carrington event not observed in most ice core nitrate records, *Geop[hysical Research Letters* **39**, L08503