On the Generative Process of Solar Flares: Non-Poisson Behavior

Abstract

The number of solar flares occurring in the corona is strongly correlated with the phase of the solar cycle. It is common practice to describe the yearly flare count distributions with a Poisson distribution. We find that the observed distributions are overdispersed relative to that expected from Poisson, and thus conclude that a Poisson generative model is not appropriate to fit to flare data aggregated in that manner. We propose that only those flares that occur within a given active region should be modeled as a Poisson process, finding that this is only the case for about 50% of active regions from which a considerable number of flares originate. The accumulation of flares from several concurrent active regions explains the observed properties of flares counts. This result has a limiting impact on assumptions for describing the physical processes of solar flare occurrences, as well as the analysis and modeling of the distribution of flares energies, which are known to be distributed as power-laws.

1 Introduction

Billions of years ago in one of the spiraling arms of the Milky Way galaxy, a dense cloud of dust and gas began to collapse under gravity, forming a protostar that spewed jets of gas into interstellar space [3]. Eventually, the amount of material pulled into this formation caused the gravity to become so intense that hydrogen atoms at its center began to fuse into helium atoms. At this point, a new star was born: the Sun [13]. Since the formation of Earth and the evolution of terrestrial life, humans have relied on the Sun to bring life to crops, harness energy for a multitude of activities, and have wondered about its origins and role in the universe.

While the sun is a necessity for humanity to thrive on Earth, it also poses a threat to the prosperity of life. The outer solar atmosphere, the corona, extends millions of kilometers into interplanetary space [7]. This non-uniform region about the sun is not easily observed with the naked eye, but can be viewed at different wavelengths of the electromagnetic spectrum or during a solar eclipse as shown in *Figure 1* [9]. Here, the sun's magnetic fields become complex, twisting and suddenly changing. These changes produce violent, extremely energetic solar storm events that permeate throughout the solar system. These events can damage satellites in orbit, disrupt communications on Earth, and in the age of space travel, can harm or even kill astronauts.

Figure 1: The corona of the sun viewed during the 2017 total eclipse. Credit: NASA

1.1 Solar Flares

The most intense of these solar storm events are solar flares, the focus of this paper. Solar flares are bursts of light and radiation caused by an impulsive release of stored magnetic energy from the sun. They are seemingly random events, whose intensities and energy releases vary over several orders of magnitude and follow power-laws [6]. Flares tend to occur in active regions, areas in the corona with strong magnetic fields that are closely associated with sunspots (see *Figure 2*).

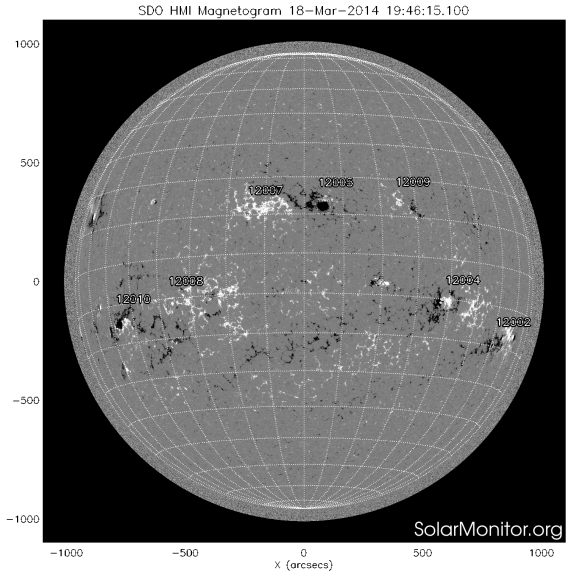

Figure 2: Several concurrent active regions (dark and light regions) on the disk of sun in March 2014. Credit: SolarMonitor.org

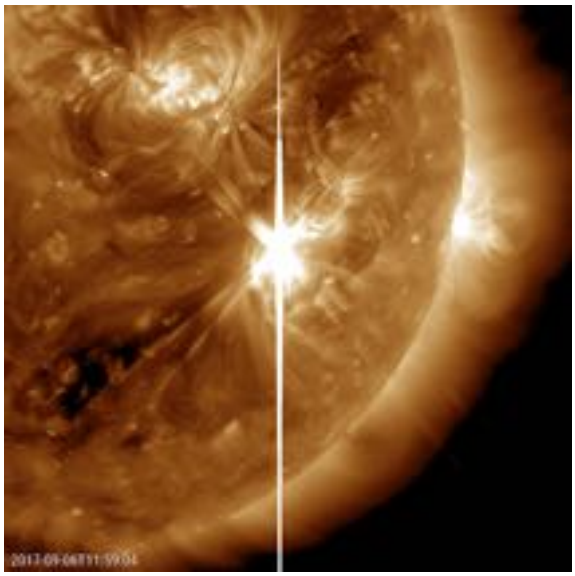

Figure 3: X9.3 class solar flare that occurred in 2017. Credit: SDO

The number of active regions changes throughout a roughly 11 year cycle of increasing and decreasing activity, called the solar cycle, because of the structural change of the sun's magnetic field. Near the solar maximum there are many active regions, and near the solar minimum there are very few. Since flares typically originate from active regions, it follows that as the number of active regions increases, then the number of flares will also increase. As a result, the frequency of flares also closely follows the solar cycle.

High intensity flares that could have adverse effects to space operations, like the one pictured in *Figure 3*, increase in frequency near the solar maximum. Thus, understanding the underlying physical processes of solar flares is in the best interest of space agencies, anyone operating or using satellites and their services, and humanity in general. Properly modeling the generative process of solar flare occurrences is an essential step to understanding and predicting flares.

1.2 Overview

Within this paper we aim to determine if solar flares follow a Poisson process using both temporal and spatial groupings, and discuss the implications of the results of our analysis. In investigating flare occurrences, it is important to acknowledge the instruments used to make flare observations and present their limitations. In addition, an in-depth understanding of the Poisson process and its assumptions are necessary before performing any statistical analysis. *Section 2* gives an overview of the satellites that are used to collect solar flare data, consider equipment and data limitations, and present the data wrangling process. Next, *Section 3* contains an in-depth explanation of the Poisson process in the context of solar flares. In *Section 4*, we outline the methods utilized to evaluate goodness of fit and also conduct a power analysis. *Section 5* covers the results of fitting a Poisson distribution to yearly 10-day flare count distributions, as well fitting an exponential distribution to the waiting times of flares within distinct active regions. Finally, the discussion and implications of the results are presented in *Section 6*.

2 Solar Flare Data

The data utilized in our research was retrieved from the Geostationary Observational Environmental Satellite (GOES) database. There has been a total of 17 GOES satellites that have been in orbit as early as 1975, all maintained by the National Aeronautics and Spaceflight Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA) [8]. Our set of data has its earliest observation of a solar flare in July of 1996, with the most recent flare in the dataset being observed in December of 2019. These observations span across solar cycles 23 and 24. The observations were made by several satellites over the 24 years of data collection, with GOES-7 and GOES-16 being the oldest and newest satellites to collect data, respectively. GOES-1 through GOES-17 are outfitted with the X-Ray Sensor (XRS), an instrument that observes the sun's soft X-ray irradiance in the 0.5-4 Angstrom (0.05-0.4 nm) and

1-8 Angstrom (0.1-0.8 nm) bands. The XRS is used to detect solar events like flares.

2.1 Flare Detection

An important part of analyzing the data from the GOES satellites is to understand how exactly a flare is detected. One may "visually" observe a flare in the corona going off by using equipment such as the Solar X-Ray Imager (SXI) on board the satellites, but it is more useful to examine the X-ray flux to track solar activity and identify flares. As previously mentioned, the X-Ray Sensors on the GOES satellites are utilized to collect data on the solar X-ray flux in the 0.5-4 Angstrom (short) and 1-8 Angstrom (long) bands at all times. The GOES long band (1-8 Angstrom) is used to detect the flares in our dataset.

The primary GOES satellite transmits 1-minute X-ray flux data back to earth in both the short and long bands. The live data can be found on the Space Weather Prediction Center (SWPC) website. *Figure 4* is a screenshot from the SWPC website of X-ray flux observed by GOES-16 over 7 days, where the red line represents the 1-minute X-ray flux in the long band and the blue line represents the 1-minute flux in the short band. It can be seen that there is a consistent "noise" in the flux in the long band, sometimes broken up by intermittent spikes. The noise is given by the constant radiation of the sun, while these significant spikes in the flux are what would typically be characterized as solar flares.

Solar flares are automatically detected by an algorithm developed for the GOES satellite. To avoid delving into the specifics of the detection algorithm, we can describe the occurrence of a flare as being determined by a peak in the flux that is significantly above the regular flux background noise. When we detect these flares we report several observed properties. These properties include the peak flux and peak time of the flare, the start and end times of the flare, and the total energy in *ergs* released by the flare. It is important to clarify how the start and peak times of flares are defined, as this is integral to the analysis being conducted. The peak time of a flare is defined as the minute at which the peak X-ray flux occurs, while the start time of the flare is defined as the first minute in a sequence of 4 minutes of steep monotonic increase in the long band flux [12].

Some flare peaks are harder to make out from the background noise when flares are small, especially when multiple flares occur in quick succession. Zooming in to a 5-hour time interval of flux data, *Figure 5* displays three solar flares that would likely be automatically detected by the algorithm. Observation *A.* points out two flares that occur in quick succession. The first flare would likely be detected, as well as the second flare. However, the rise time of the first flare is clear while the second flare's rise time is not. In addition, it is difficult to determine the descent of the first flare since these two flare fluxes overlap. Observation *B.* is a single flare that can easily be made out from the background noise, having a clear rise, peak, and descent in flux.

2.2 Data Wrangling

Once a flare is observed through the processes described in the previous sections, its measured properties are recorded and placed in a database. We formulate our dataset using data originating from the GOES flare database, only utilizing observations recorded by the GOES-7 through GOES-16 satellites between 1996 and 2019. This time period covers almost the entirety of solar cycles 23 and 24. The flares' start and end times, peak times and fluxes, and total energies are not the only properties recorded. We also collect information on their locations (longitude and latitude), the active regions in which they occurred, data quality of the observations, and several more properties. For the purposes of this research it is not necessary to use all the included measurements, so the utilized measurements are given in *Table 1* alongside their corresponding definitions and units.

Before utilizing the solar flare data in our

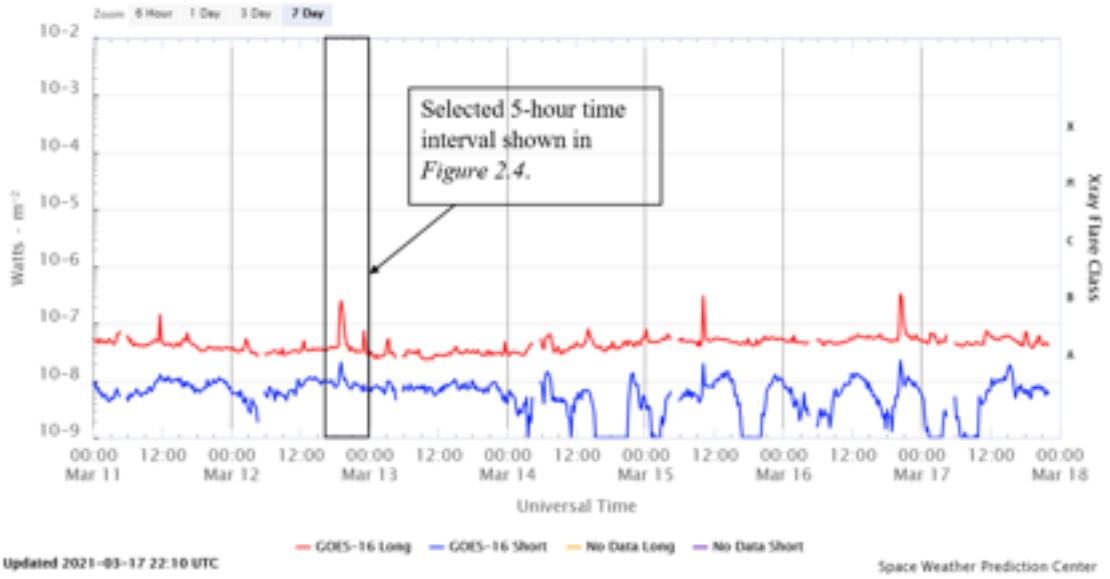

Figure 4: Example of X-ray flux observed by GOES-16 (March 11-18, 2021). The red line is the 1-minute X-ray flux in the long band that we use to determine the occurrence of a solar flare. Credit: SWPC

Figure 5: Flares are detected by peaks in the flux above the regular background flux noise in the long band (red) (March 12, 2021). Observation *A.* identifies two B-class flares that occur in quick succession, making it difficult to collect accurate data. Observation *B.* is a single A-class flare that is easily identified. Credit: SWPC

Variable	Definition
Peak Flux	peak flux in long band at earth ($ergs/s/cm^2$)
Total Energy	total energy released at sun ($ergs$)
Start Time	start time in YYYY-MM-DDTHH:MM:SS
Peak Time	time of peak flux in YYYY-MM-DDTHH:MM:SS
Duration	time the flare lasted (sec)
Active Region #	active region number assigned to flare, if known
Longitude	longitude of location flare occurred, if known
Latitude	latitude of location flare occurred, if known
GOES Satellite	GOES satellite that observed the flare

Table 1: List of variables utilized in our dataset.

analysis, it is necessary to properly prepare the data. The pre-processed dataset has a total of 38,114 observations. We outright remove any observations that are designated as NA. If an observation has an unreasonable value due to a recording error for the variables given in *Table 1*, such as a negative total energy value, these observations are also outright removed. Although a small proportion of observations, we filter out observations that are designated as poor quality by the GOES satellites or not made in the 1-8 Angstrom band. Once these observations are removed, making up roughly 12.25% of the observations in the original data, the processed data has a total of 33,445 observations.

3 Solar Flares as a Poisson Process

The number of flares occurring at any given time is dependent on the phase of the solar cycle. We observe that the frequency of flares increases as the solar maximum is approached, and decreases near the solar minimum. This trend can be viewed in the aggregated flare counts through the bar plot shown in *Figure 6*, where each bar represents one year.

This trend may be self-evident with even a limited understanding of solar physics, but figuring out how the flare counts are distributed does not have nearly as obvious of an answer. Flares can be thought of as random events that occur at some rate. This rate is non-stationary through time given the cyclical nature of flare counts. However, if we select some interval of time such that the rate is stationary, then the flare counts could follow a Poisson distribution and their occurrences modeled by a Poisson process.

3.1 Poisson Process

The occurrence of flares are random, meaning the precise time at which a flare occurs is unknown, but we may assume that they occur at some constant rate λ. Let $N(t)$ be the number of flares that occurred at or before time t, such that $N(0) = 0$. If we observe the sun for a time where $t \in [0, \infty)$, we may break this time up into tiny intervals of length δ, such that from $[0, t]$ there are $m = t/\delta$ intervals. For each interval we may observe at most one flare, so we can treat the occurrence of a flare in each interval as a Bernoulli trial with probability $p = \lambda\delta$ of a flare occurring. Further assuming that the occurrences of flares in disjoint time intervals are independent, $N(t)$ follows a Binomial distribution

$$P(N(t) = n) = \binom{m}{n} p^n (1-p)^{m-n}. \qquad (1)$$

We know that as $\delta \to 0$, $m \to \infty$ and $p \to 0$ the Binomial distribution approaches the Poisson distribution. The parameter for this Pois-

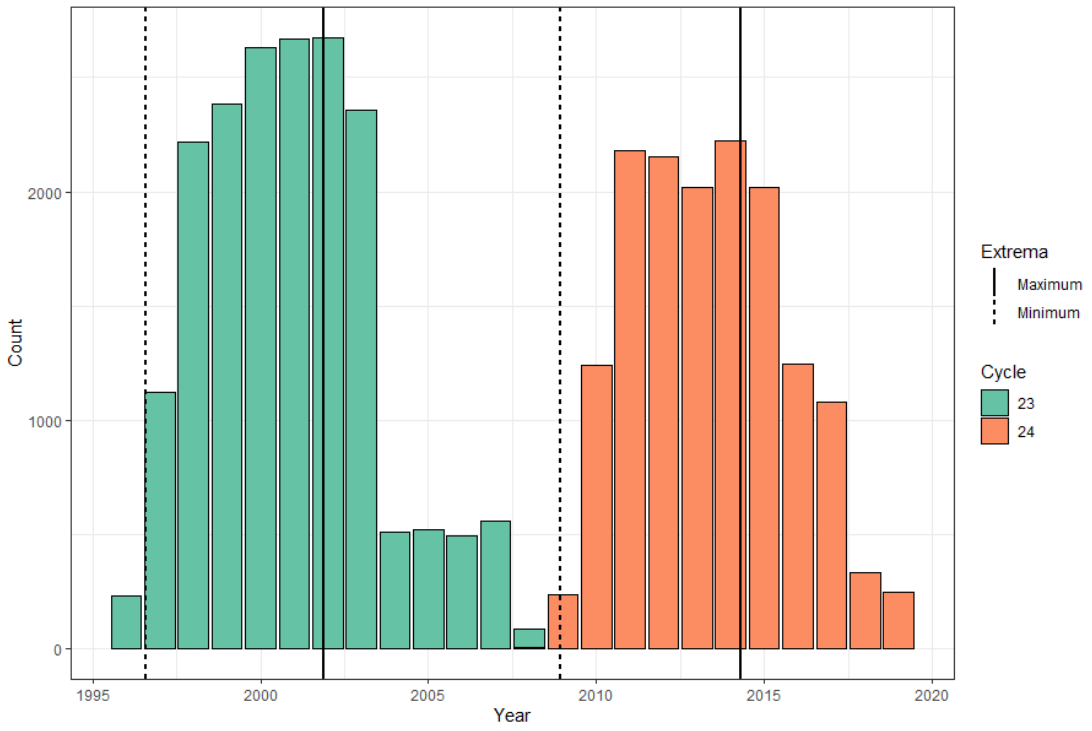

Figure 6: Bar plot of aggregated flare counts per year throughout solar cycles 23 and 24, where each bar represents one year.

son distribution is $mp = \frac{t}{\delta}\lambda\delta = \lambda t$, so $N(t)$ follows a Poisson distribution

$$P(N(t) = n) = \frac{e^{-(\lambda t)}(\lambda t)^n}{n!}. \tag{2}$$

What has been outlined is known as a Poisson process. It is of our interest to model flare occurrences using a Poisson process, if it is appropriate, because of its properties and implications for understanding the physics of flares. We may select any length of time and the number of flares occurring should follow a Poisson distribution with some rate λ of flares per unit time. If we choose to model flares this way, we make the following assumptions,

1. the occurrences of a flare in one time interval is independent of the occurrence of a flare in any other time interval;
2. for a given length of time, the rate of occurrence λ is constant.

As a consequence, the time between two successive flare occurrences, what we call the waiting time, follows an exponential distribution. To see this, let W_i be the time elapsed between the $i-1$ and ith flare (called the "waiting time"), then for the first flare

$$\begin{aligned} P(W_1 > t) &= P(N(t) = 0) \\ &= \frac{e^{-(\lambda t)}(\lambda t)^0}{0!} \\ &= e^{-(\lambda t)}. \end{aligned} \tag{3}$$

The cdf $F_{W_1}(t) = P(W_1 \leq t) = 1 - P(W_1 > t)$, such that

$$F_{W_1}(t) = \begin{cases} 1 - e^{-\lambda t}, & \text{if } t \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This is the cdf for an exponential distribution. Now let W_2 be the waiting time between the first and second flare. If the first flare occurs at time s and t is the elapsed time between the first and second flare, where $s, t > 0$, then it follows that the second flare does not occur

in the interval $(s, s+t]$. Alternatively, we may write

$$P(W_2 > t|W_1 = s) = \\ P(\text{no flare occurs in } (s, s+t]|W_1 = s). \tag{5}$$

Since the intervals $(0, s]$ and $(s, s+t]$ are disjoint, flare occurrences in these intervals are independent and

$$\begin{aligned} P(W_2 > t|W_1 = s) &= P(W_2 > t) \\ &= P(\text{no flare occurs in } (s, s+t]) \\ &= e^{-\lambda t}. \end{aligned} \tag{6}$$

The cdf for W_2 is that of an exponential distribution, so the waiting time W_2 follows an exponential distribution with parameter λ. It can then be shown using the independence assumption that for any W_i, where W_i is the waiting time between the $i-1$ and ith flare, $W_i \sim Exponential(\lambda)$. Thus, if the flare counts follow a Poisson distribution, the waiting times will follow an exponential distribution, and flare occurrences can be modeled by a Poisson process. If the waiting times between flares follow an exponential distribution, does this imply the flare counts follow a Poisson distribution? Answering this is not trivial, but it can be shown to be true through the following.

If W_i denotes the waiting time between the $i-1$ and ith flare, where $W_i \sim Exponential(\lambda)$, let T_n be the time at which the nth flare occurs and

$$T_n = \sum_{i=1}^{n} W_i \text{ where } n \geq 1. \tag{7}$$

The sum of the n exponential distributions results in $T_n \sim gamma(n, \lambda)$. The pdf of T_n is given by

$$f_{T_n}(x) = \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} \text{ for } x \geq 0 \tag{8}$$

Notice that for $n = 1$, $T_1 \sim Exponential(\lambda)$ [5].

Now, to demonstrate that $N(t) \sim Poisson(\lambda t)$, we follow the proof for Theorem 9.1.1 by Ross 2019 [15]. Consider that the number of flares that occur at or before time t is at least n if and only if the nth flare occurs at or before time t, written as

$$N(t) \geq n \Leftrightarrow T_n \leq t. \tag{9}$$

This implies that

$$\begin{aligned} &P(N(t) = n) = \\ &P(N(t) \geq n) - P(N(t) \geq n+1) \\ &= P(T_n \leq t) - P(T_{n+1} \leq t) \\ &= \int_0^t \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} dx \\ &\quad - \int_0^t \lambda e^{-\lambda x} \frac{(\lambda x)^n}{n!} dx. \end{aligned} \tag{10}$$

This can be simplified by rearranging the integration by parts formula, $\int udv = uv - \int vdu$, to be $uv = \int udv + \int vdu$. If $u = e^{-\lambda x}$ and $dv = \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} dx$, this results in

$$\begin{aligned} e^{-\lambda t} \frac{(\lambda t)^n}{n!} &= \int_0^t \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} dx \\ &\quad - \int_0^t \lambda e^{-\lambda x} \frac{(\lambda x)^n}{n!} dx, \end{aligned} \tag{11}$$

completing the proof.

The preceding work shows that if the waiting times between flares follow an exponential distribution, their sum, the time for the nth flare to occur, will follow a gamma distribution. The number of flares n that occur at or before time t, $N(t)$, then follows a Poisson distribution. This will be important to know in the following sections on modeling flare occurrences through a Poisson process.

4 Evaluating Goodness of Fit

In attempting to model flare occurrences using a Poisson process, we must verify that flare counts follow a Poisson distribution or that their waiting times follow an exponential distribution. We do this using

probability-probability (P-P) plots on a case-by-case basis to visualize the fits, while also utilizing simulation-based versions of both the Kolmogorov-Smirnov (KS) test and the Anderson-Darling (AD) test when working with large numbers of fitted distributions.

4.1 Lilliefors Test

Perhaps one of the most well known and easily implementable goodness of fit tests, the Kolmogorov-Smirnov (K-S) test statistic measures the absolute value of the maximum distance between the theoretical cdf F and the empirical cdf F_n. It is used to test the goodness of fit of a distribution to sample data and the test statistic is given by

$$D = \max_x |F_n(x) - F(x)|. \tag{12}$$

In conducting the K-S test, we assume that the observed data follows a given theoretical (null) distribution under the null hypothesis. We set our significance level to $\alpha = 0.05$ and if we obtain a p-value below this threshold, we reject the null hypothesis and have significant evidence to suggest that the observed data does not follow the distribution defined under the null hypothesis. The validity of this test is dependent on how the null distribution is defined. Often, in empirical studies, the parameters of the null distribution cannot be specified a priori but need to be estimated from the sample. This estimation-based specification of the null distribution is used to conduct the K-S test. As a consequence, the distribution of the K-S test statistic changes and the way the p-value was computed can no longer be used [11].

Since we estimate the parameters from our sample, we must employ alternative methods to properly utilize the K-S test. We implement the Lilliefors test, which is a simulation-based method of the K-S test when defining the null distribution with estimated parameters. The Lilliefors test has been shown to have more power than the regular K-S test, though it is more computationally demanding. The Lilliefors test is performed as follows:

1. Estimate the parameters $\hat{\theta}_0$ of the given distribution from the observed sample of size n.
2. Calculate the K-S statistic D_0 whose null distribution F_0 is evaluated at the estimated parameters $\hat{\theta}_0$.
3. Draw a random sample of size n from F_0.
4. Estimate the parameters $\hat{\theta}_i$ from this new random sample.
5. Calculate the K-S statistic D_i whose null distribution F_i is defined by the estimated parameters $\hat{\theta}_i$.
6. Perform steps 3. through 5. 10,000 times $(i = 1, \ldots, 10,000)$.
7. Calculate the proportion of $D_i \geq D_0$. This is the new p-value for the K-S statistic D_0.

If the new p-value is less than the significance level α, we have significant evidence to suggest that the observed data does not follow the null distribution. We choose to implement the Lilliefors test over the regular K-S test in our analysis because it has more power and the correct type I error rate. See *Section 4.3* for a power analysis.

4.2 Corrected Anderson-Darling (A-D) Test

The Anderson-Darling (A-D) test is a test used to determine the goodness of fit of a given distribution to a sample. In testing for normality, the A-D test has been shown to be a powerful test and has similar power to that of the Shapiro-Wilks test [14]. The test statistic is a modification of the Cramér-von Mises (CVM) test, where the weight function in the A-D test gives more weight to the observed values in the tails of the distribution. The A-D test statistic is given by

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x), \tag{13}$$

where F is the specified theoretical (null) cdf, F_n is the empirical cdf, and $[F(x)(1 - F(x))]^{-1}$ is the weight function. Under the null hypothesis for this test, we assume that our sample came from the specified null distribution. If the p-value is below our significance level, $\alpha = 0.05$, then we reject the null hypothesis and have sufficient evidence to suggest that the null distribution is not a good fit to the observed data. Caution is advised when using this test, just as with the K-S test. The distribution of the test statistic changes when the null distribution is specified by estimated parameters from the sample. Braun proposed a method for conducting the K-S test that accounts for estimated parameters, but it requires large sample sizes and we show it performs poorly through a power analysis [4].

Instead, we implement the same steps of the Lilliefors test given in *Section 4.1* but with the Anderson-Darling test statistic. After conducting a power analysis, we have shown that the Lilliefors-based method of conducting the A-D test has the highest power. We choose to rely on this version of the A-D test[1] as our primary goodness of fit test, with the Lilliefors test being a sanity check. Implementing both of these tests alongside P-P plots gives us a robust way of evaluating the goodness of fits of the Poisson and exponential distributions to the observed solar flare count and waiting time distributions, respectively.

4.3 Power Analysis

We are interested in utilizing the goodness of fit test with the most power when testing for an exponential distribution. To do so, we perform a power analysis of the Kolmogorov-Smirnov test, the Lilliefors test, Anderson-Darling test, the Braun-Adjusted Anderson-Darling test, and the corrected Anderson-Darling test. In conducting the power analysis, we simulate data drawn from several distributions, then fit an exponential distribution to the simulated data. We then examine the performance of each test to determine their power. After considering the simulation-based methods, we conduct over 2.7 billion simulations.

[1] We refer to this as the "corrected A-D test" for the remainder of the paper.

4.3.1 Goodness of Fit Tests

The Kolmogorov-Smirnov test and the Anderson-Darling test are two well known goodness of fit tests that are easily implementable in various situations. However, it is necessary to test their power before relying on their results for inference, especially when defining the theoretical (null) distribution using estimated parameters from a sample. We outline the K-S test, A-D test, and corresponding modifications to these tests in this section.

Komlogorov-Smirnov (K-S) Test The K-S test statistic measures the absolute value of the maximum distance between the theoretical cdf F and the empirical cdf F_n. The test statistic is given by

$$D = \max_x |F_n(x) - F(x)|. \tag{14}$$

The null and alternative hypotheses of the K-S test are given by,

H_0 : Sample follows null distribution

H_a : Sample does not follow null distribution

If D is above the threshold of the $1 - \alpha$ quantile, where α is the designated significance level, then we reject the null hypothesis.

Lilliefors Test When the parameters of the null distribution are specified from a sample, the distribution of the K-S statistic changes and the way the p-value is regularly computed can no longer be used. The Lilliefors is a simulation-based method of the K-S test when defining the null distribution with estimated parameters. Under the null hypothesis, we assume that the sample follows the null distribution. If the new p-value we calculate is below α, then we reject the null hypothesis. The

steps taken to perform the Lilliefors test are outlined in *Section 4.1*.

Anderson-Darling (A-D) Test The Anderson-Darling (A-D) test is a modification of the Cramér–von Mises (CVM) test, where the weight function in the A-D test gives more weight to the observed values of the distribution. The A-D test statistic is given by

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x), \quad (15)$$

where F is the specified theoretical (null) cdf, F_n is the empirical cdf, and $[F(x)(1 - F(x))]^{-1}$ is the weight function. Under the null hypothesis for this test, we assume that our sample came from the specified null distribution. If A^2 exceeds the $1 - \alpha$ quantile, then we reject the null hypothesis.

Braun-Adjusted A-D Test The A-D test is subject to the same issues as the K-S test when the null distribution is specified by estimated parameters from a sample. When conducting the regular A-D test, we utilize the *ad.test* function from the *goftest* package in R. This function allows the user to indicate if the parameters provided for the null distribution are estimated from the sample. If this is indicated, then the function implements the Braun-adjusted A-D test to adjust for estimated parameters. We choose to not outline the methods of conducting this test in this paper, but the null and alternative hypotheses are the same as the A-D test.

Corrected A-D Test An alternative to the Braun-Adjusted A-D test is used a simulation-based version of the A-D test. We conduct this corrected A-D test in the same way as how the Lilliefors test is conducted, outlined in *Section 4.1*, except we implement the A-D test statistic instead. Under the null hypothesis, we assume that the sample follows the null distribution. If the new p-value we calculate is below α, then we reject the null hypothesis.

4.4 Procedures

We use Monte Carlo methods to evaluate the power of the K-S test, the Lilliefors test, A-D test, the Braun-Adjusted A-D test, and the corrected A-D test for an exponential distribution in this power analysis. We set the level of significance α to 0.05 for all the tests we conduct. The null and alternative hypotheses of the tests are given by

H_0 : Sample follows null distribution
H_a : Sample does not follow null distribution.

To determine the power of each test, we simulate data drawn from 4 distributions for 17 sample sizes ranging from $n = 10$ to $n = 200$. The four distributions are: $Exponential(\lambda = 0.5)$, $Weibull(k = 1.4, \lambda = 2.5)$, $Lognormal(\mu = 0.1, \sigma^2 = 0.6)$, and $Gamma(a = 2, s = 0.5)$. For each distribution and sample size, we draw a total of 10,000 samples and conduct calculate all the test statistics for all simulated samples. Including the simulations in the Lilliefors and corrected A-D test, we perform the tests on 2.7 billion simulations. To save an immense amount of time and better utilize available computational power, we implement parallel processing using the *parallel* package in R and the run time is only about 27 hours.

4.5 Power Analysis Results

The power of each test is determined by sample size and the distribution in which the simulated sample was drawn from. At all sample sizes, the power of each test is higher when the simulated sample was drawn from a distribution that differs more from the exponential distribution. We can determine the type-I error rate of the tests when $\alpha = 0.05$ by finding the percent of tests that result in a rejection of the null hypothesis for samples drawn from the exponential distribution. As seen in *Figure* **??**, the type-I error rate of the Lilliefors and corrected A-D test is approximately 5%

for all sample sizes. The regular K-S and A-D tests are highly conservative, committing type-I errors less than 1% of the time for all sample sizes. The Braun-Adjusted A-D test approaches about 3.75% as the sample size increases, so it is also conservative.

For samples drawn from non-exponential distributions, we find that the corrected Anderson-Darling test has the highest power at all sample sizes, closely followed by the Lilliefors test. The Braun-Adjusted A-D test performs very poorly, having a power of about 2% power at all sample sizes. The K-S test and the A-D test are very close in power for all sample sizes. However, the K-S test tends to have more power at low sample sizes, eventually being taken over by the A-D test. The results of the tests for an exponential distribution for samples drawn from the 4 different distributions are given in *Figure 7*, *Figure 8*, *Figure 9*, and *Figure 10*.

As previously mentioned, the power of each test is higher for all sample sizes when the simulated sample was drawn from a distribution that differs more from the exponential distribution at all sample sizes. For example, the lognormal distribution we draw from differs the most from the exponential distribution, so the tests have high power at even lower sample sizes. Alternatively, the Weibull distribution we draw from is the closest to the exponential distribution, so the tests only gain more power at higher sample sizes. We conclude that the corrected Anderson-Darling test has the highest power, closely followed by the Lilliefors test, both being robust enough for use at sample sizes of about about 50.

5 Results

5.1 Distribution of Flare Counts By Year

Historically, solar physicists have fit a Poisson distribution to solar flare count data over different time intervals, often making the assumption that it constitutes a good fit. We investigate this in depth because if a Poisson distribution is appropriate, then we are able to model flare occurrences through a Poisson process. If we can model flare occurrences this way, then we may take full advantage of the many nice properties of a Poisson process and use the physical framework of self-organized criticality for solar flares [2]. However, one of the major assumptions of a Poisson process is that the rate of occurrences is stable, which is evidently not the case as seen in *Figure 11*. If the Poisson distribution is not a good fit, it has several implications on modeling approaches and on understanding the physics of solar flares.

Our approach to fitting to the GOES flare occurrence data is to treat each year as its own dataset. Using by-year data is common practice for astronomers, but the assumption that the rate parameter must be stable throughout each year is doubtful. Admittedly, the time unit of a year is completely arbitrary relative to the dynamics of the sun, but it is a nice unit that humans are familiar with. To fit the Poisson distribution to by-year data, we find the number of flares that occur every ten days within each year, then fit a Poisson distribution to the distribution of 10-day counts within each year. Other recent approaches to studying flare occurrences, specifically their waiting times, use a non-stationary Poisson process, which allows for a continuous change in the rate of occurrence parameter λ throughout solar cycles [10][1]. These approaches are more sophisticated and take into account the unstable rate of occurrence, but do not consider the grouping of flares that we take into consider later in this chapter. Nevertheless, using by-year data is an easily interpreted and common approach taken by solar physicists to study the dynamics of the sun over time.

5.1.1 Overdispersion

Using the GOES data from 1997 to 2019, we break the 365 calendar days for each year into $k = 37$ bins of length 10 (e.g. $(0, 10], \ldots, (360, 370]$). For each year, we count the number of flares that occurred within all of the 10-day periods. We then filter out a total

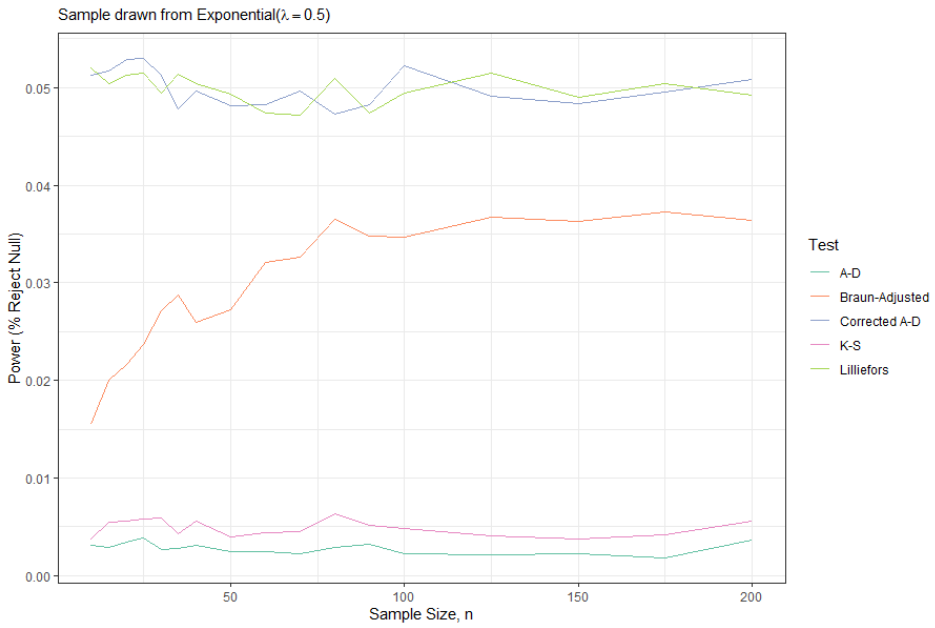

Figure 7: Power of tests for exponential for the simulated samples from an exponential distribution. The power in this case is the type-I error rate, where $\alpha = 0.05$. The corrected A-D test and the Lilliefors have the expected error rate of 5%, whereas the other tests are conservative.

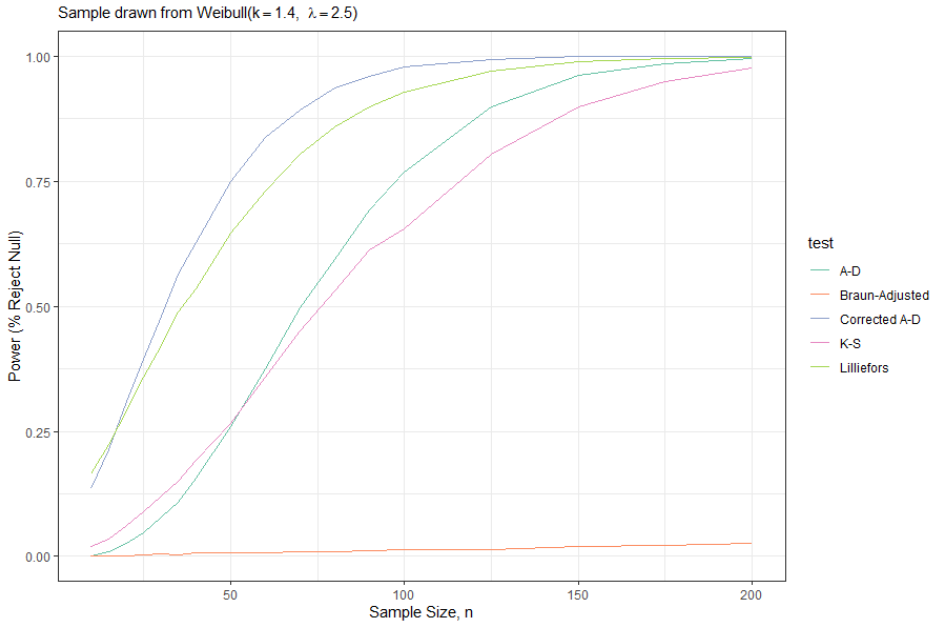

Figure 8: Power of tests for exponential for the simulated samples from an Weibull distribution. The Weibull distribution differs the least from the exponential distribution, so the power for all tests does not get sufficiently high until larger sample sizes. The corrected A-D test has the highest power in this case, having considerable power at a sample size of 50.

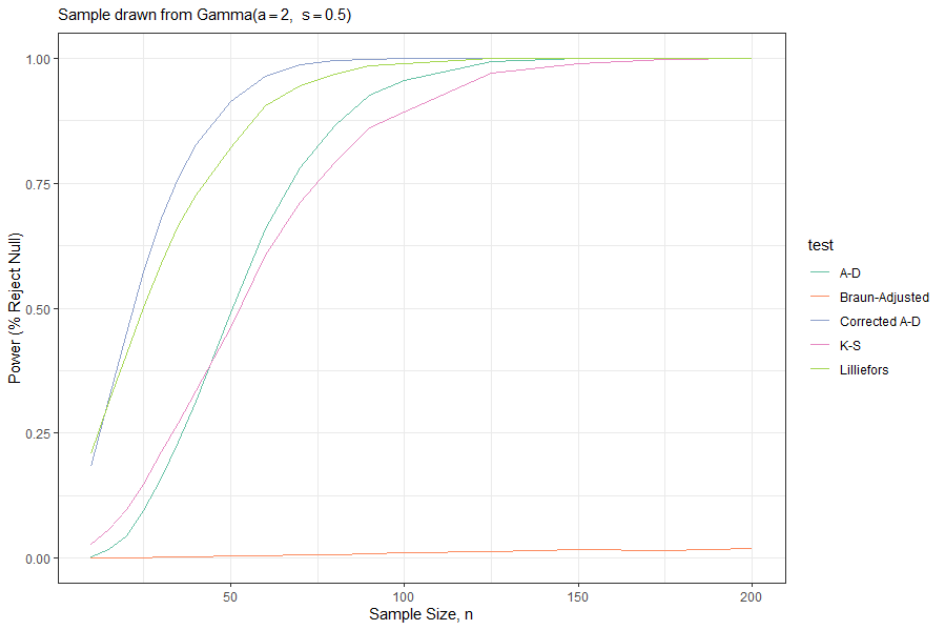

Figure 9: Power of tests for exponential for the simulated samples from a gamma distribution. The corrected A-D test has the highest power in this case, having considerable power at a sample size of 40. The Braun-Adjusted A-D test performs terribly, being much too conservative.

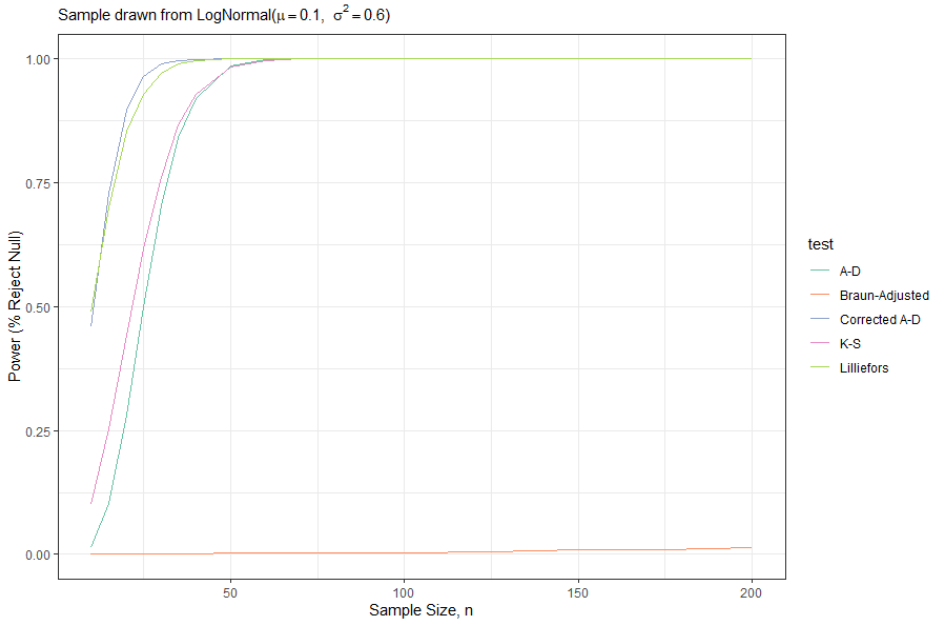

Figure 10: Power of tests for exponential for the simulated samples from a lognormal distribution. The lognormal distribution differs the most from the exponential distribution, so the power of all tests quickly increases as sample size increases. The corrected A-D test has the highest power in this case, having considerable power at a sample size of 15.

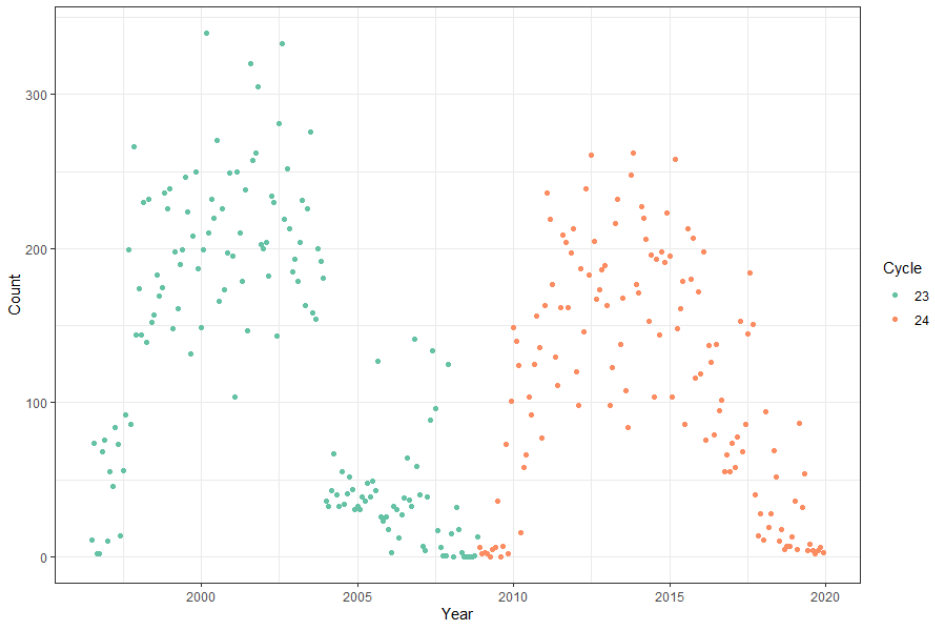

Figure 11: Monthly flare counts throughout solar cycles 23 and 24.

of 17 observations that occurred on February 29 of the leap years to keep the bins consistent. In addition, we remove any flares that occurred in the year 1996 because we only have data starting from the end of July 1996.

Let a random variable N_j be the number of flares that occur every 10 days in year j, $j = 1997, \ldots, 2019$. If N_j follows a Poisson distribution then its pmf is given by

$$P_{N_j}(N_j = n_j) = \frac{e^{-\lambda_j}\lambda_j^{n_j}}{n_j!}, \quad (16)$$

where $n_j = 0, 1, 2, \ldots$ and $\lambda_j > 0$.

Here, n_j is the number of flares that occurred in a 10-day period in year j and λ_j is the rate of occurrence (flares per 10 days) for year j. If N_j follows a Poisson distribution, then the occurrences of flares in year j follow a Poisson process.

We fit the Poisson distribution to the 10-day count data for each year j using the maximum likelihood estimate for the rate of occurrence λ_j. The MLE is given by the average of the 37 10-day counts in year j, $\hat{\lambda_j} = \frac{1}{37}\sum_{i=1}^{37} x_i$. As seen in *Figure 11*, the observed 10-day counts vary considerably throughout the solar cycles, even within a given year. The result is an overdispersed observed distribution relative to what is expected under Poisson. This overdispersion is also demonstrated in the P-P plots in *Figure 12*, where the points tend to dip below the identity line and do not stay within the confidence bands. To further prove the poor goodness of fit of the Poisson distribution to this data, we conduct the corrected A-D test and Lilliefors test[2] We find that the the Poisson distribution is not a reasonable fit for 100% and 97% of years according to the corrected A-D test and Lilliefors test, respectively.

The results of both the visual interpretation and formal tests convincingly show that the Poisson distribution is not a reasonable fit to by-year 10-day flare count data. If binnings other than 10-day bins are used for counts, or even other time intervals besides years, the Poisson distribution is still not a reasonable fit and the observed distributions are still overdispersed. This suggests that flare occurrences

[2]Note that we do not conduct a power analysis for testing for the goodness of fit of a Poisson distribution. However, it is quite obvious that the Poisson is not reasonable in almost all cases anyhow.

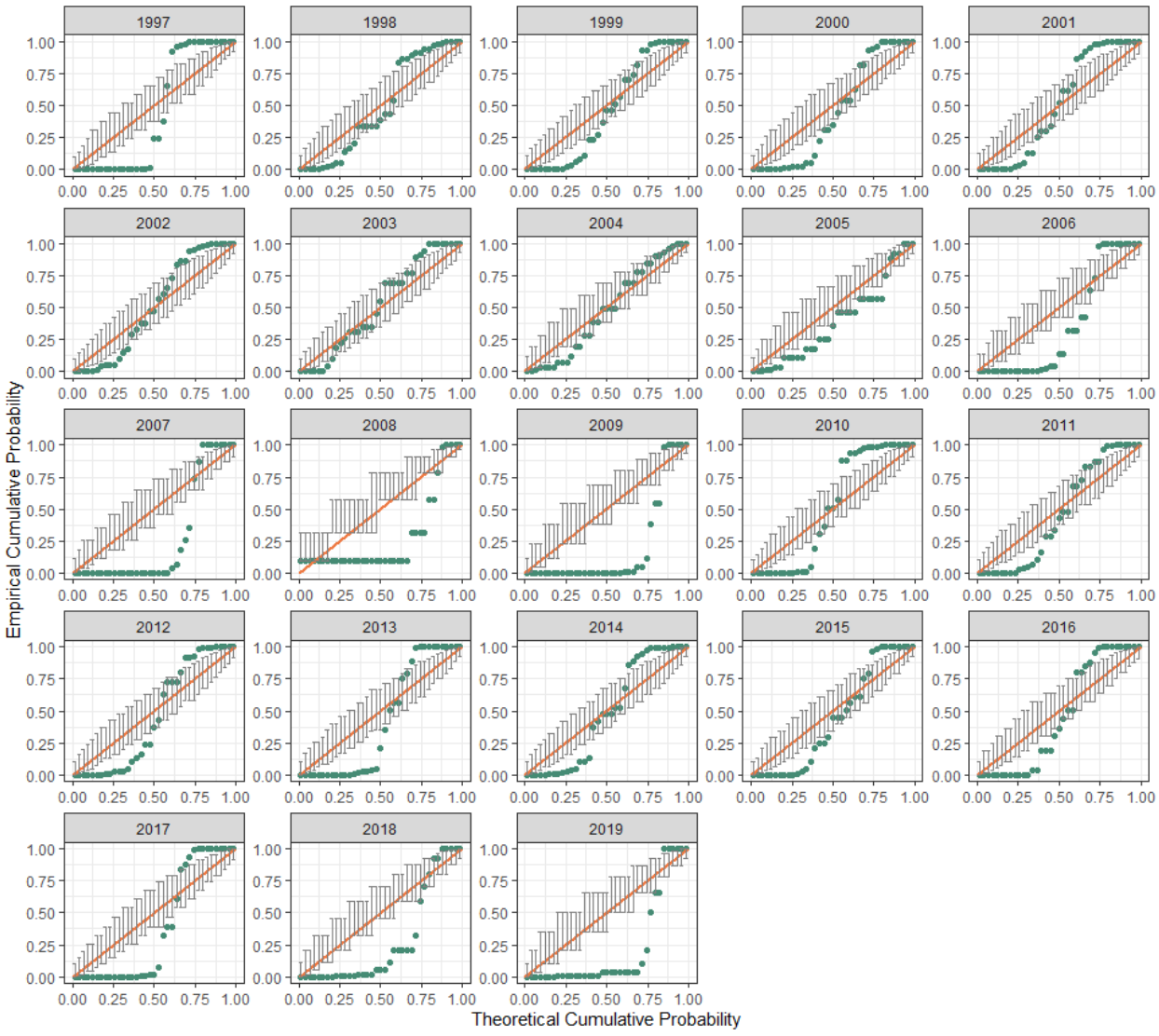

Figure 12: P-P plot of the fitted Poisson distributions for each year. We observe that many of the points for each year are outside the confidence bands and tend to dip below the 45° line, so the observed distribution within each year is overdispersed.

cannot be modeled by a Poisson process. The overdispersion is a result of the solar cycle, which quite clearly has a relationship with the number of flares occurring. However, before ruling out a Poisson process it is important to consider the fact that flares tend to occur in spatial groupings on the sun, specifically in active regions.

5.2 Solar Flare Occurrences Within Active Regions

Active regions are places where the magnetic fields emerge from inside the sun, which is the result of the process of magnetic field lines "entangling" throughout the solar cycle [7]. The structure of the magnetic field changes throughout the cycle, resulting in the number of active regions changing. Since flares tend to occur in these regions, it follows that as the number of active regions increases, then the number of flares increases, and vice versa. Investigating flare occurrences within active regions, rather than on an aggregate level, may be the key to understanding the generative processes of solar flares.

Active regions are unique in several ways, so they should not be treated as identical. Spatially, active regions differ in their size and their latitudes on the sun[3]. Active regions are also unique in their temporal characteristics, as they occur at different times and last for different lengths of times. Additionally, the number of observed flares within active regions varies to a great extent.

The qualities of these active regions may differ, but the physical laws that govern their dynamics remain the same. Thus, the process of flare occurrences should be similar throughout active regions, so studying each active region individually can lead to insights into how flares occur and may explain the overdispersion in flare counts observed in the aggregate. We attempt to determine the appropriateness of a Poisson process for modeling flares within active regions. Since the length of time active regions last differ, fitting a Poisson distribution to count data within active regions is difficult. Instead, we fit an exponential distribution to the waiting times within active regions. As proven in *Section 3.1*, if the waiting times between flares follow an exponential distribution, then flares can be modeled through a Poisson process.

5.2.1 Distribution of Waiting Times

We define the waiting time as the difference in time between the start times of two consecutive flares. We choose to use the starting time of flares in defining the waiting times because the interpretation is intuitive, consistent with the Poisson process, and the accuracy of start times measurements is reasonable, even when compared to using the peak times of flares. The accuracy of the measured start times of flares is within about one minute, except when flares occur at close times. In this case, the start time of two flares near each other in time is harder to determine, but only 14.5% of waiting times are at or below one minute for all observed flares, including those not assigned to an active region, so this should not pose a major issue.

Let the random variable W_r, $r = 1, ..., 1518$, be the waiting time between two consecutive flares within the same active region r. If W_r follows an exponential distribution, then its pdf is given by

$$f_{W_r}(W_r = w_r) = \begin{cases} \lambda_r e^{-\lambda_r w_r}, & \text{if } w_r \geq 0 \\ 0, & w_r < 0. \end{cases} \tag{17}$$

Here, λ_r, $\lambda_r > 0$, is the rate parameter for active region r and can be interpreted in the same way as that of the Poisson distribution, except here we define it as the number of flares per hour. To fit the exponential distribution to the observed waiting times for each active region we find $\hat{\lambda}_r$ by maximum likelihood estimation. For the exponential distribution it

[3]We recommend looking up "The Butterfly Diagram" in relation to active regions, solar flares, and sun spots. It is a beautiful graphical visualization of a physical phenomenon.

is given by $\hat{\lambda}_r = \frac{1}{\bar{w}_r}$, where $\bar{w_r}$ is the mean waiting time within active region r.

We fit the exponential distribution to all 1,518 active regions in the GOES database, but are only interested in regions with a sufficiently large number of flares. After conducting a power analysis of the Lilliefors test and corrected A-D test, we find that a sample size of 50 results in the tests having enough power to be reliable. For this reason, we choose to focus on active regions with at least 50 flares. There are only 52 active regions with at least 50 flares, which is quite a small number. We also choose to fit to active regions with at least 30 flares, increasing the number of active regions to 142. There is a loss of power of the tests at that sample size, so the results for active regions with less than 50 flares should be considered with caution.

After conducting both tests on the distributions of waiting times for active regions, we find that an exponential distribution is a reasonable fit for the distribution of waiting times in 48% of active regions with at least 50 flares according to the corrected Anderson-Darling test. By the Lilliefors test, the exponential distribution is a reasonable fit for 52% of active regions. The numbers do increase when we consider active regions with 30 or more flares. The exponential distribution is a reasonable fit for 51% of active regions by the correct A-D test and 58% by the Lilliefors test. The results are summarized in *Table 2*.

To investigate the distributions of the waiting times of some active regions, we select 20 active regions with between 30 and 35 flares. The P-P plots for the exponential fits for these active regions are given in *Figure 13*. There are definitely some active regions where the exponential distribution constitutes a very good fit, but others where it is a very poor fit. It seems that the observed distributions of the waiting times tend to be overdispersed, not underdispersed, for the cases where the exponential distribution is not a reasonable fit.

We investigate relationships between several characteristics of active regions and whether or not an exponential distribution was a good fit for that region. We find that there is not relationship with the goodness of fit with the number of flares in an active region, the rate parameter, and the average latitude of the active region. In addition, we observe no sign of a temporal relationship or relationship with the total energy output of active regions.

6 Conclusion

6.1 Discussion of Results

The results of our analysis on the 10-day counts for each year throughout solar cycles 23 and 24 show that a Poisson distribution is not a reasonable fit, as there is too much variation in the counts between and within years. After investigating the distribution of waiting times within active regions, the results suggest all active regions cannot be modeled through a Poisson process. However, there are about 50% of active regions where the flares can be modeled through a Poisson process. This mixture is interesting, as it could suggest that there are physical processes that change the way in which flares occur in different active regions. One major assumption of the Poisson process is that events are independent, so we must assume this about flares within active regions. When a flare goes off the structure of the magnetic fields in the area are altered, which could have an effect on other flares going off in the local corona [7]. This could be one reason for the mixed results for the by-active region fits.

Whether or not we assume that the flares within active regions follow a Poisson process, we are able to explain the overdispersion seen in the 10-day count data. This overdispersion is caused by the random accumulation of flares from multiple concurrent active regions at any given point in time. Suppose flare occurrences within these active regions follow a Poisson process, then their counts will follow a Poisson distribution. If we sum these distributions, the resulting distribution will also be Poisson distribution with a new rate parameter. So for any time segment the aggregated count data should follow a Poisson distribu-

Test	Minimum # Flares	# Active Regions	% retained
Corrected A-D	50	52	48%
Lilliefors	50	52	52%
Corrected A-D	30	144	51%
Lilliefors	30	144	57%

Table 2: Results of correct A-D and Lilliefors tests for exponential distribution of waiting times within active regions.

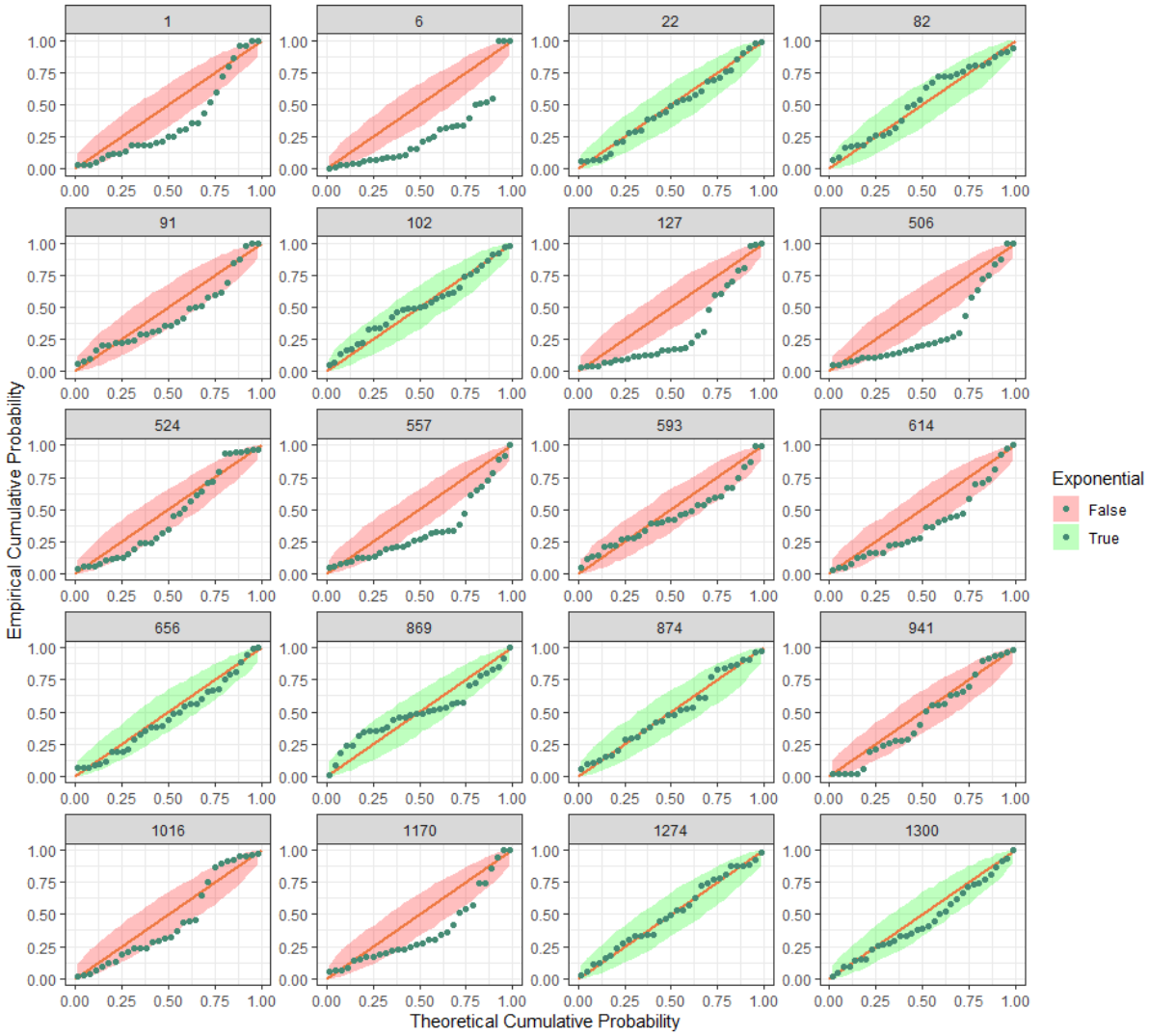

Figure 13: P-P plot of the fitted exponential distributions to flare waiting times within 20 active regions with 30 to 35 flares. For some of the plots, the points line up well to the 45° line, while for others the distributions are overdispersed. The red confidence bands indicate that the corrected A-D test gave evidence to suggest that the exponential distribution was a poor fit.

tion. However, this is only when the active regions occurring together are always the same. Instead, active regions appear and dissipate after differing lengths of times, with the total number appearing increasing with the solar cycle. Because of this, there is a random number of active regions at any given moment, each producing its own flares at its own rate.

Over the course of 10 days, there might be 5 different active regions that produced flares, but for only 5 of those days 3 were active and the other days the other 2 were active. Although the flare counts within active regions may be distributed as Poisson, the accumulation over these 10 days results in a changing rate parameter because of the changing number of active regions. This idea can be seen in the the observed 10-day count data, shown in *Figure 14*. The graph plots the latitude of the active regions that occurred in 2013, with the length of each line segment giving the time for which the active regions were active. The color of the line segments are given by the log of the rate parameter for each active region. Several times throughout 2013 the active regions overlap, which would result in changing rates if their counts all followed a Poisson distribution.

While the results suggest that flare occurrences do not follow a Poisson process for about half of the active regions, the reason for overdisperson in the aggregated flare count data presented still reveals insight into why the number of flares occurring vary so much. We choose to make the assumption that flares follow a Poisson process for the purposes of modeling flare energy distributions, so this should be taken into account when considering the analysis to be performed later on. A model that allows for more variation, such as the Weibull or Negative Binomial distributions, might be more appropriate for flare data. These results have major implications for the area of solar physics research because result of using a non-Poisson process to model flares could entirely change the understanding of the physics of solar flares.

6.2 Limitations

Given the fact that the equipment and algorithm are not perfect at detecting flares, it is expected that there are some limitations to our data. We identify the major limitations in four specific forms. The first limitation is one of missing data resulting from the location of our satellites. Due to the satellites being in orbit around earth, they only observe one side of the sun's disk at any given point in time. This results in the satellites not recording flares that occur beyond the limbs of the observed disk. This loss of data is consistent.

The second is caused by sensitivity limitations of the X-Ray Sensors on the GOES satellites. While the equipment has become more sensitive with newer iterations, the sensors are not perfect at detecting the change in flux across all magnitudes. This makes it difficult to observe X-ray events that we may otherwise define as flares because they have such a low flux. The result is many undetected flares at low fluxes. The third limitation occurs when flares of higher fluxes increase in frequency around the solar maximum, resulting in low flux flares being harder to differentiate from the flux originating from the high flux flares. In a similar fashion, the fourth limitation is an issue of detecting flares that occur in quick succession. When flares occur close together, it may be hard to make out the peaks of all the flares. This also makes it difficult to obtain accurate estimates of the start and end times of the flares. These four limitations are some of the contributing factors to the struggles of collecting accurate data on solar flares, especially when attempting to detect low flux flares.

6.3 Moving Forward

The way in which we decide to model flare occurrences is important in understanding the generative process of flares, their underlying physics, and in developing models for the distribution of flare properties, such as their total energy. These properties tend to follow power-

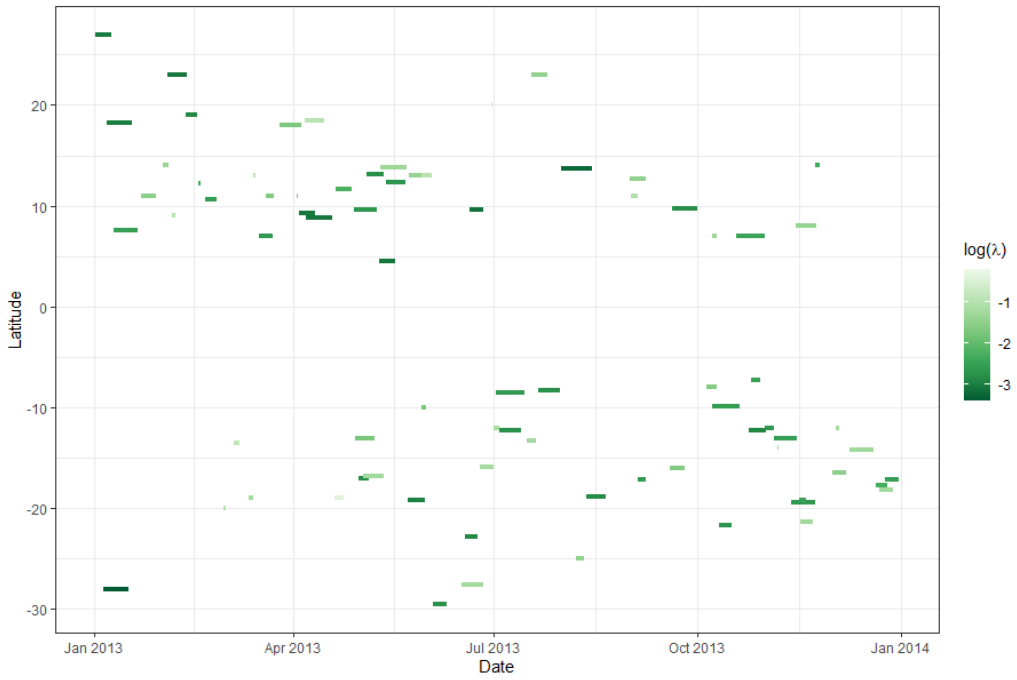

Figure 14: The existence of several active regions lap at any given moment, each with their own rate parameter. The length of each segment represents the time for which the active region lasted, with the color corresponding to the log of that region's rate parameter λ

laws and we use several statistical methods, such as Maximum Likelihood and the Maximum Product of Spacings methods, to fit this distribution to the flare data.

Moving forward, we hope to

- Conduct an even more robust power analysis.
- Investigate new ways to model flare occurrences (besides through a Poisson process).
- Apply these results to modeling of power-law distributions for solar flare properties
- Further develop a maximum likelihood method to fit a power-law to flare energy distributions.

References

[1] Markus J. Aschwanden and James M. McTiernan. "Reconciliation of Waiting Time Statistics of Solar Flares Observed in Hard X-rays". In: *The Astrophysics Journal* 717.2 (July 2010), pp. 683–692. DOI: 10.1088/0004-637X/717/2/683. arXiv: 1002.4869 [astro-ph.SR].

[2] Per Bak, Chao Tang, and Kurt Wiesenfeld. "Self-organized criticality: An explanation of the 1/fnoise". In: *Physical Review Letters* 59.4 (1987), pp. 381–384. DOI: 10.1103/physrevlett.59.381.

[3] Jeffrey O. Bennett et al. *The Essential Cosmic Perspective*. Pearson, 2009.

[4] Henry Braun. "A Simple Method for Testing Goodness of Fit in the Presence of Nuisance Parameters". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.1 (1980), pp. 53–63. DOI: 10.1111/j.2517-6161.1980.tb01100.x.

[5] George Casella and Roger L. Berger. *Statistical inference*. Duxbury, 2002.

[6] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4 (Jan. 2009), pp. 661–703. DOI: 10.1137/070710111.

[7] Leon Golub and Jay M. Pasachoff. *The Solar Corona*. 2009.

[8] Lynn Jenner. *GOES Overview and History*. Mar. 2015. URL: https://www.nasa.gov/content/goes-overview/index.html.

[9] M. Kretzschmar. "The Sun as a star: observations of white-light flares". In: *Astronomy & Astrophysics* 530, A84 (June 2011), A84. DOI: 10.1051/0004-6361/201015930. arXiv: 1103.3125 [astro-ph.SR].

[10] C. Li et al. "Waiting time distributions of solar and stellar flares: Poisson process or with memory?" In: *Monthly Notices of the Royal Astronomical Society: Letters* 479.1 (Sept. 2018), pp. L139–L142. DOI: 10.1093/mnrasl/sly117.

[11] Hubert W. Lilliefors. "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown". In: *Journal of the American Statistical Association* 64.325 (1969), pp. 387–389. DOI: 10.1080/01621459.1969.10500983.

[12] NASA and NOAA. *GOES X-ray Flux*. Mar. 2020. URL: https://www.swpc.noaa.gov/products/goes-x-ray-flux.

[13] *Our Solar System*. Dec. 2019. URL: https://solarsystem.nasa.gov/solar-system/our-solar-system/in-depth/.

[14] Nornadiah Mohd Razali. "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling Tests". In: *Journal of Statistical Modeling and Analytics* 2.1 (Jan. 2011), pp. 21–33.

[15] Sheldon M Ross. *A first course in probability*. Sixth edition. Upper Saddle River, N.J. : Prentice Hall, 2002. URL: https://search.library.wisc.edu/catalog/999921659402121.