Proposal to the

Smithsonian Astrophysical Observatory (SAO)

for work on the

Statistical Analysis, Prediction,

Modeling, and Visualization

of Solar Flare Data

in conjunction with the grant proposal

“Many Flares Make a Corona”

Submitted by

Bernhard Klingenberg

Dept. of Mathematics & Statistics

Williams College, MA

June 9, 2023

**1. Description and Outline of Work**

**1.1 Overview**

The primary work will be on the development, implementation and visualization of statistical procedures to model characteristics of solar flares based on satellite observations of hundreds of thousands of flares spanning two and a half solar cycles. A particular focus will be the development of a probabilistic model for the energy release for such flares. This is complicated by the fact that flares with low energy output are harder to observe than those with large energy output. The signature of low-energy flares is harder to detect amid the background noise, but also during times when the sun is very active and signatures of the many high energy flares swamp the signals from the lower energy flares.

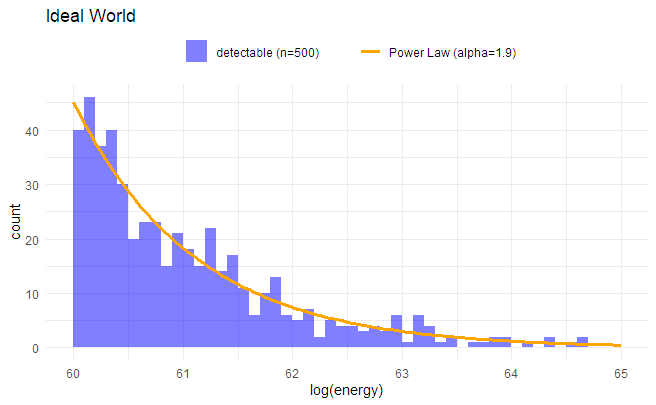
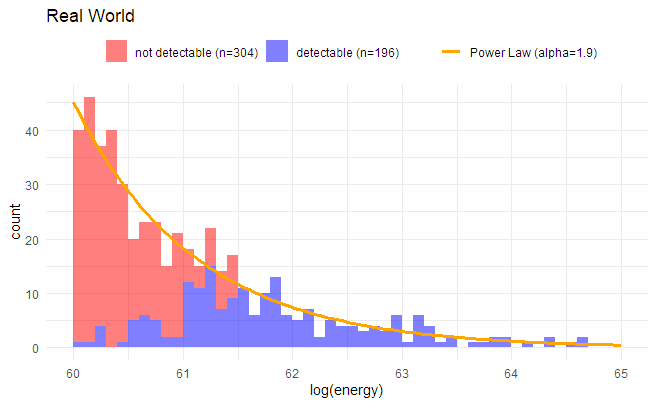
Figure 1 illustrates the issue using simulated data. In an ideal world we could observe all flares occurring on the (visible part of the) sun, and, based on astrophysical considerations, might assume a Power-Law for the distribution of their energy releases, see the left panel in Figure 1. However, because of instrument and measurement limitations, many flares that have lower energy output can’t be identified from analyzing the stream of flux data, and are therefore missing from the data. Consequently, the energy distribution that we actually observe looks more like the blue distribution in the right panel of Figure 1.

Figure 1: Distribution of flare energies in an ideal world without measurement limitations (left panel) and with measurement limitations (right panel), where many low-energy flares cannot be observed.

Before exploring various statistical methods (maximum likelihood, maximum product of spacings, Bayesian) towards building a stochastic model for the flare energies based on the existing data on solar-flare energies, we hope to significantly supplement and enrich the existing data by identifying flares that have been missed using conventional algorithms of flare detections. We plan to use machine learning methods, in particular a convolutional neural network (CNN) or similar deep learning algorithms, to predict where flares may have occurred based on the available flux data. Similar methodology was used successfully by Feinstein et al. (2020)[[1]](#footnote-1) to identify flares in transient exoplanet survey satellite short-cadence data, using an existing hand-labeled catalog of such flares to train the CNN.

**1.2 Augmenting the existing flare catalog**

The database that we are using for our analysis consists of 1-minute averages of X-Ray irradiance measurements from the GOES satellite[[2]](#footnote-2). This database includes information on when a potential flare occurred, determined from an algorithm based on the relative magnitude and rise of the flux measurements[[3]](#footnote-3). Our plan is to use this existing source of identified flares to train our CNN. Note that our goal is not, as in many other projects, forecasting when the next flare will happen based on observations in the past, but rather, and perhaps simpler, to automatically identify flares in a large and existing dataset.

To illustrate and test our research proposal, we created a dataset of all 1-minuted fluxes for the years 2021 and 2022 from the GOES 16 catalog that is available online[[4]](#footnote-4). From it, we compiled a dataset of 4,203 flares that were identified in those two years. Similar to the approach taken by Feinstein et al. (2020), for each of these flares, we constructed a flare profile that consists of the fluxes 30 minutes prior to the flux for which the flare event was declared (called the peak flux), and the fluxes for the 15 minutes after the peak flux. Thereby, a flare profile consists of 46 flux values (with the peak flux in position 31), and covers a period of 46 minutes. Due to missing data, we could not create a profile for all 4,203 flares, but for 4,180 of them.

*Research Goal:* These values for a time span of 46 minutes were picked heuristically, and research needs to be done to assess if such a time span is sufficient for the CNN to learn about the background flux against which to assess the probable occurrence of a flare.

We did not use the raw flux measurements directly (they are measured in W/m2), but divided each of the 46 flux values in a profile by the flux value in position 31, which is where the peak flux was observed. Consequently, the value in position 31 of any profile is always equal to 1, while the other values are typically less than 1 (when there were no other flares in the 46-minute window). However, they can also be higher than one if the initial flux in a profile was higher (e.g., on a downward trend from a previous flare), or if another flare occurred before or right after the one considered. Figure 2 (top row) shows four selected flare profiles that are included in our dataset. These profiles will be used as the positive cases (labeled ‘flare’) in the training dataset.

To create a negative case (or control) profile, we randomly selected a minute in a given year, but excluding those minutes where a flare was observed (i.e., where we had a positive cases). We also excluded the 3 minutes immediately before or after the identified flare. For a randomly selected minute outside those windows, we created the flare profile by combining the 30 flux values before the selected minute, at the selected minute, and the 15 flux values after the selected minute. This means that the control profile also shows flux values for a span of 46 minutes, but where flux values in positions 28 to 34 are known to not be at or very near a flaring event. We used the same standardization technique (dividing by the flux in position 31) as for the flare profiles. Figure 2 (bottom row) shows four such selected control profiles. They were labeled as ‘no flare’ in the training dataset.

In total, we created control profiles in a ratio of 30:1: For every identified flare in a given year, we randomly selected 30 control profile. Overall, we thus have 4,180 + 125,400 = 129,580 profiles in our dataset.

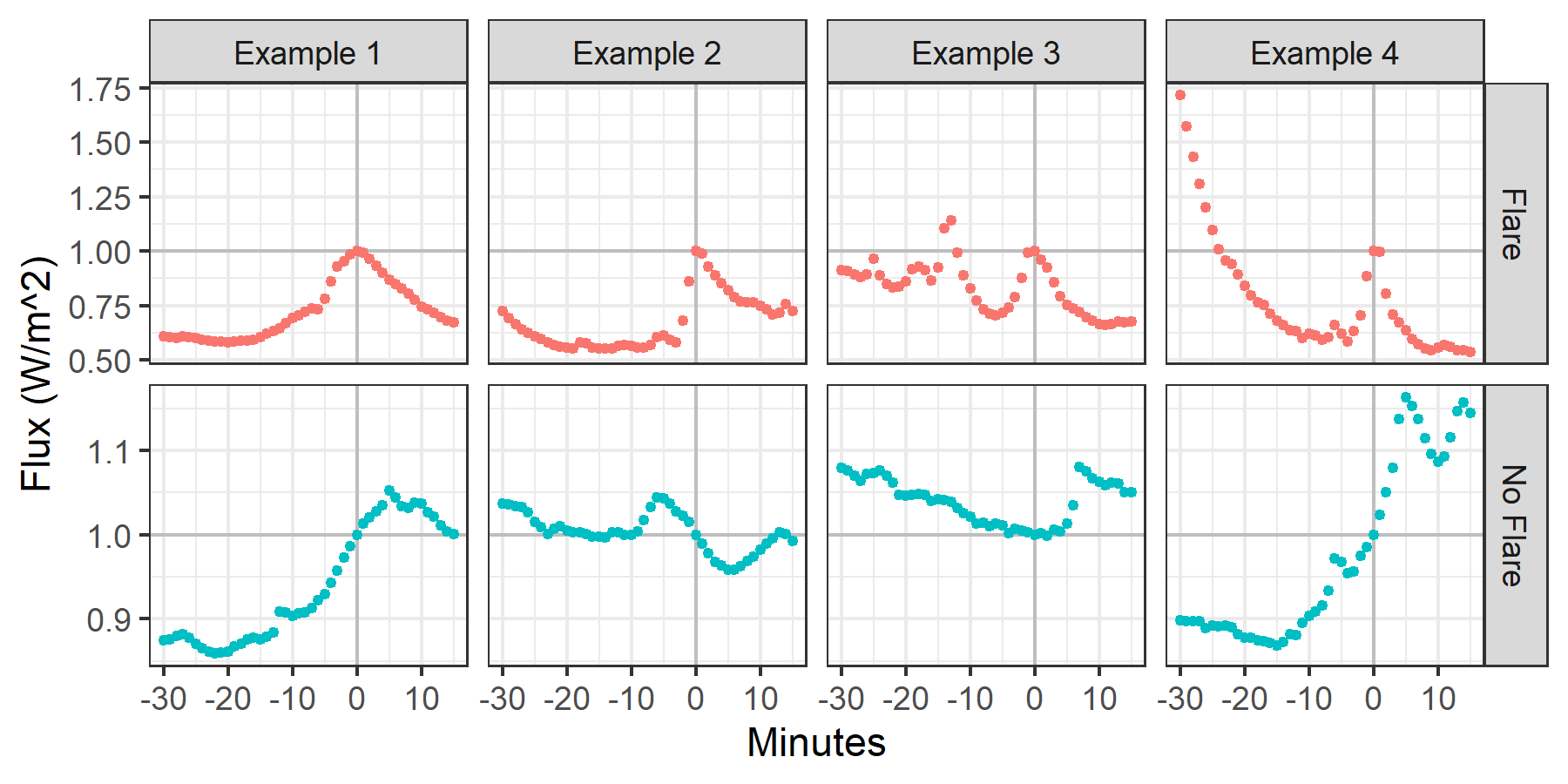


Figure 2: Examples from the training dataset of various profiles of flares (top row) and non-flares, the controls (bottom row). Profiles consist of 46 consecutive flux values, each divided by the flux value in position 31. We labeled position 31 as minute 0, the first flux observation in a profile as minute -30 (as it occurs 30 minutes before), and the last observation as minute 15.

We used 80% of this labeled (‘flare’, ‘no flare’) profile dataset to train the CNN, holding out the remaining 20% for evaluation. For the architecture of the CNN, for this proposal, we used the architecture and hyperparameters of the *stella* CNN developed in Feinstein et al. (2020), with minimal adjustments.

*Research Goal:* Considerable effort will have to be devoted to develop, train and optimize a CNN that is sensitive to the characteristics of the GOES X-ray data. Feinstein et al. (2020) specifically mention their CNN should only be used for light curves as derived from the transiting exoplanet survey satellite (TESS) data. Feinstein et al. (2020) published their CNN algorithm in a github repository[[5]](#footnote-5), but they use an older Python library that now is obsolete, and we could not reproduce their analysis. We plan to published all data and code necessary to train and evaluate the CNN that is developed under this research in its own github repository, so other researchers can use it for flare analysis of GOES X-ray data.

Using the training data as described above (which contained 3,356 flare profiles) to train the CNN, we obtained an accuracy of 96.49% (= 24,263/25,145), a precision of only 48.26% (= 804/1,666) and a recall of 97.57% (= 804/824) on the test data, which contained 824 flares. The high recall is desirable, as the CNN should be able to reproduce the flares that the conventional method identified. The low precision is not of concern at this stage, as we know the current flare catalog missed many of the smaller flares, which we hope our CNN will pick up. We plan to investigate significant more time in obtaining and optimizing training datasets, and fine-tuning the CNN. This will lead to improved metrics.

To illustrate the potential of our approach for expanding the existing flare catalog, Figure 3 shows per-minute flux measurements taken between February 2 and February 11, 2022. During this period, 76 time points were identified as the peak flux of a flare, using the conventional algorithm. These time points are published on the GOES website and are shown as red dots in Figure 3.

To obtain a prediction of the likelihood of a flare going off at any particular minute, we created a flux profile as described above: We put the flux at the desired minute in position 31 of the profile, with positions 1 – 30 consisting of the fluxes of the previous 30 minutes, and positions 32 – 46 consisting of the fluxes of the following 15 minutes. This profile is then fed into our CNN and we obtain a prediction score as an output. This prediction score can be used to predict whether the input profile at that minute has the characteristics of a flare. We did this for every single minute in the time interval. The green dots in Figure 3 indicate all time points where the CNN predicted the profile was that of a flare. (Additional research is needed into settings for hyperparameters, most importantly the threshold over which a profile is declared a flare. Here, we used 1/30, the ratio of flares to non-flares in our training dataset.).

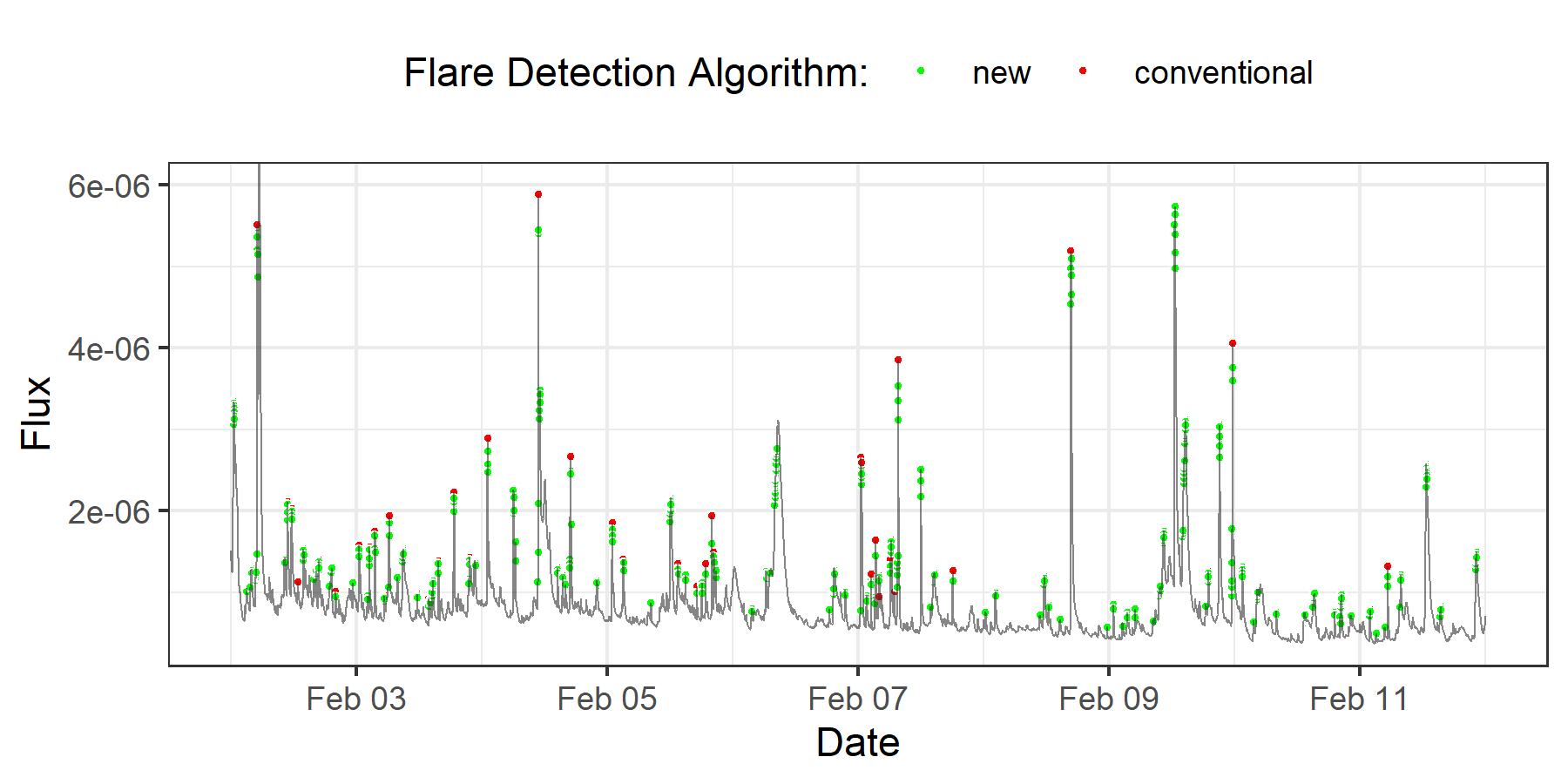


Figure 3: Fluxes observed between February 2 and February 11, 2022 (gray line). Red dots indicate time points (unit = 1 minute) where the conventional algorithm indicated a flare. In all but 1 of the 76 cases a flare was identified by the conventional algorithm, it was also identified by the CNN. Green dots indicate additional time points where the CNN identified a flare, but the conventional method did not.

Using our very preliminary CNN, it is clear that we are able to identify many more flares, in particular those of smaller amplitudes. And judging by the flow of the flux, the overwhelming majority of those newly identified time points indeed seem to correspond to flares. Once we research how to obtain the energy release of these newly identified flares, we have a much richer database to model the energy distribution.

**1.3 Modeling the flare energy distribution**

With a sufficiently supplemented dataset of flares, our focus will now shift towards modeling their energy release. This entails fitting various statistical models to the available data, which traditionally has been done using maximum likelihood methods. These methods, however, have known limitations in this setting: Because the energy output of flares, being sudden bursts generated by the magnetic field, varies over several order of magnitudes, it may be reasonable to model the energy release as a power law (see Figure 1). Estimation of the power law index (alpha) is complicated by the fact that the onset of the power law (the energy level below which no flares would occur) is unknown and maximum likelihood estimators for it may not perform satisfactorily. A significant amount of my work will investigate a multinomial approach, based on binning the data, and a Bayesian approach to model fitting. Preliminary simulation results carried out by me look very promising in more accurately estimating these key characteristics of flare energy distributions.

Another strand of work will relate to investigate the almost canonical assumption that the number of flares with energies in a given (small) interval follows a Poisson distribution. Our preliminary analysis show that this assumption, for various reasons mentioned elsewhere, may not be valid. I plan to investigate alternative models that take over-dispersion relative to the Poisson model into account. The Bayesian approach is especially suited for such an approach, as an extension of the Poisson distribution is the negative binomial distribution, resulting from the mixture of a Poisson distribution with a gamma distribution. My hope is that such a modeling approach more accurately predicts the average number of flares occurring, but also accurately describes the variability in the number of flares.

A parametric modeling approach as outlined so far is not the only approach towards analyzing solar flare characteristics. Another part of the research in this grant relates to a novel semi-parametric approach to model fitting, with model fitting carried out by the product of maximum spacings algorithm. Part of my work will be a comparison between these two different approaches. It will be interesting to see whether the more traditional parametric approach results in similar conclusions and estimates as the semi-parametric approach.

Finally, for both model development and dissemination of our results to the larger research community, it is important to create appropriate visuals that help in the description of the model and our predictions, but also visualize uncertainties. Part of my work will focus on creating online, interactive web apps based on the work described here.

**2. Expertise**

I’m a Professor of Statistics and recently served as the director of a graduate program in Data Science. I have extensive training in categorical response data and specializing in statistical modeling, consulting and data visualization. I will provide my statistical expertise, including expertise in the statistical computer language R and the R shiny package for creating interactive web apps to the overall goal of understanding and modeling key characteristic of solar flares.

**3. Time**

I plan to work for 1.5 months for each of three years to carry out the work given in the description above.

|  |  |
| --- | --- |
| **Year** | **Description of Work** |
| 1 | Obtaining a large, labeled (‘flare’, ‘no flare’) training dataset from published GOES X-ray fluxes and perhaps some other sources. Using the training dataset to develop a Convolutional Neural Network for flare identification. |
| 2 | Fine-tuning the CNN model and supplementing the existing GOES flare catalog with newly identified flares from the CNN. Obtaining the total energy output for the newly identified flares. Explore Maximum Likelihood and Bayesian model fitting of the energy distribution. |
| 3 | Writing up results and developing visuals for internal and external dissemination of results, including the creation of web apps. Creating a Github repository that makes the obtained expanded flare catalog public and shows reproducible code for all method used in the research, including the training and application of the trained CNN. |

**4. Travel**

Although most research meetings with the larger team can be carried out over Zoom, the collaborative nature of this research with experts in astrophysics and statistics necessitates a few in-person meetings to effectively discuss modeling and implementation issues, and to obtain and incorporate feedback that is too difficult to relay over Zoom. Since the principal investigator is located in Cambridge, MA, I’m asking for travel between Williamstown, MA and Cambridge, MA on three different occasions per year.

For each year, I’m budgeting for three trips to Cambridge, each for three days (two nights).

For the second year, I’m budgeting a three-day, two-night trip to the University of Michigan to meet with grant collaborators to work on the parametric modeling approach to solar flare analysis.

This work is interesting to the broader statistical community and in year two I’m planning to present our approach and current results at the joint statistical meeting, held in Nashville, TN, in August 2025 (year 2 of the contract).

**5. Current & Pending Support**

I have no current or pending support from any other grant or agency.

1. Feinstein, Montet, Ansdell, Nord et al. (2022). Flare Statistics for Young Stars from a Convolutional Neural Network Analysis of TESS Data. *The Astronomical Journal*, 160:219 [↑](#footnote-ref-1)
2. https://data.ngdc.noaa.gov/platforms/solar-space-observing-satellites/goes/goes16/l2/docs/GOES-R\_XRS\_L2\_Data\_Readme.pdf [↑](#footnote-ref-2)
3. See Appendix A in https://data.ngdc.noaa.gov/platforms/solar-space-observing-satellites/goes/goes16/l2/docs/GOES-R\_XRS\_L2\_Data\_Users\_Guide.pdf [↑](#footnote-ref-3)
4. https://www.ngdc.noaa.gov/stp/satellite/goes-r.html [↑](#footnote-ref-4)
5. https://github.com/afeinstein20/stella [↑](#footnote-ref-5)