The Astrophysical Journal Letters, 937:L8 (6pp), 2022 September 20 https://doi.org/10.3847/2041-8213/ac8de9

Clustering Behavior in Solar Flare Dynamics

Elmer C. Rivera[1], Jay R. Johnson[1], Jonathan Homan[1], and Simon Wing[2]
[1] Andrews University, Berrien Springs, MI, 49104-0660, USA; jrj@andrews.edu
[2] Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099, USA
Received 2022 June 17; revised 2022 August 21; accepted 2022 August 31; published 2022 September 20

Abstract

The solar magnetic activity cycle provides energy input that is released in intense bursts of radiation known as solar flares. As such, the dynamics of the activity cycle is embedded in the sequence of times between the flare events. Recent analysis shows that solar flares exhibit memory on different timescales. These previous studies showed that the time ordering of flare events is not random, but rather there is dependence between successive flares. In the present work, the clustering of flares is demonstrated through a straightforward nonparametric method where the cumulative distribution function of successive flares is compared with the cumulative distribution function of surrogate sequences of flares obtained by random permutation of flares. The random permutation is performed within rate-variable Bayesian blocks during which the flare rate is assumed to be constant. Differences between the cumulative distribution functions are substantial on a timescale around 3 hr, suggesting that flare recurrence on that timescale is more likely than would be expected if the waiting time were drawn from a nonstationary Poisson process.

Unified Astronomy Thesaurus concepts: Solar flares (1496)

1. Introduction

Substantial energy is released in the form of flares during the magnetic activity cycle of the Sun. The statistics of these flares provide valuable information about the underlying dynamics of the cycle. Flares are driven by the dynamics of active regions, which result when magnetic flux driven by the solar dynamo emerges from the interior of the Sun (Wing et al. 2018; Charbonneau 2020). As this flux emerges, the convective motions that bring the flux to the surface twist and tangle the magnetic field leading to the development of intense current sheets, which are generally thought to release their energy through magnetic reconnection and X-ray emissions from electrons accelerated during this process (Toriumi & Wang 2019, and references therein). The release of energy for these flare events generally occurs on a timescale that is short (about 20 minutes) compared with the typical time between the events around 5 hr (Snelling et al. 2020). As such, it is useful to consider flares as a sequence of discrete events, which can be characterized by the time interval, Δ_j, between the $j^{\rm th}$ and the $j+1^{\rm th}$ events.

Because of the close relationship between the flares and the solar dynamo, it is expected that the flare sequence would have information about the solar cycle. Indeed, because more solar activity drives more flares, the rate of flares changes throughout the solar cycle leading to a peak rate near solar maximum and a minimum rate near solar minimum. Wheatland & Litvinenko (2002) refer to this type of driving as "external" and explored how these changes affect the statistics of flare waiting times. On the other hand, the dynamics of individual flares occur on a much shorter timescale and may well reflect only the local dynamics of an active region. As such, it is expected that flares from separate active regions are more likely to occur independently and randomly. A number of studies have shown that the statistical distribution of flares is consistent with these concepts (e.g., Wheatland 2000b; Moon et al. 2001; Wheatland & Litvinenko 2002; Aschwanden 2019). That is, the distribution of flares is consistent with a time variable Poisson process where the probability of a flare occurring in a time Δt is simply $\lambda(t)\Delta t$, where $\lambda(t)$ is the average rate, which changes slowly as a function of time. Similar processes are also manifested in the statistics of floods and earthquakes (Hong & Guo 1995; Gilroy & McCuen 2012).

The manner in which the rate changes as a function of time has been shown to affect the asymptotic power-law exponent of the distribution (Aschwanden & McTiernan 2010; Aschwanden et al. 2021). Recently, from a parameterization of the time intervals with coherent growth in the rate of events (time structures) it was found that the power law changes when considering different timescales, revealing the dynamics of the solar dynamo, partial occultation of flare events, and clustering of flares. The power-law exponent is also believed to demonstrate the memory over timescales of a few hours to several decades, which can be attributed to clustering of solar flares and a dynamo-driven solar cycle, respectively (Aschwanden & Johnson 2021). Cyclical changes of rate lead to a theoretical power law $P(\Delta) \propto \Delta^{-2.5}$ that is consistent with the observed power law of flares (Nurhan et al. 2021). Clustering of flares has also been associated with an overabundance of short waiting times (10 s–10 minutes) compared with simulated distributions (Wheatland et al. 1998). Other studies revealed the existence of memory in the flare production in the whole solar disk (Lepreti et al. 2001) and also in an individual solar active region (Lei et al. 2020) from the evidence that the stable distribution and the power-law-tail Lévy function fit well with the waiting time distribution of the solar flares. Under the same distinctive criterion of memory capability, Li et al. (2018) found that the waiting time distribution of the weaker solar flares is a process with memory that can be described with the Weibull distribution.

While characteristics of the statistical distribution of waiting times suggest memory in flares, they may not be the best method to identify memory because such studies do not

consider the relationship between pairs or groups of flares. A more direct measure of dependence can be found in the approach of Snelling et al. (2020), who examined the mutual information between successive flares. In that study the mutual information $\mathcal{M}(\Delta_n, \Delta_{n+p})$ was considered as a function of "lookahead," p, which examined how much information is common between a flare waiting time and another flare waiting time; p steps ahead in the sequence of waiting times. The results of this analysis were compared with an identical analysis of surrogate sequences of flares constructed using Bayesian blocks (Scargle 1998; Wheatland & Litvinenko 2002). The surrogates were consistent with a nonstationary Poisson process. The analysis revealed that flares up to about $p = 6$ tend to have some relationship, but it was not possible to distinguish flares with longer "lookahead" from a nonstationary Poisson process. The p=6 lookahead corresponds to a timescale of approximately 1 day. Overall, this analysis showed fairly clearly that a short-term memory exists for flares and it provided a timescale.

Mutual information as a discriminating statistic is particularly useful because it provides a single statistic that characterizes linear and nonlinear relationships between random variables, and it has been widely applied to study nonlinear dependence in solar and magnetospheric systems (Johnson & Wing 2005, 2014; Wing et al. 2016, 2018, 2020, 2022; Johnson et al. 2018; Wing & Johnson 2019). However, it is even more interesting to explore the joint statistical distributions of flares to see the origin of the relationship that was detected. In particular, by looking at the joint distribution, we can see how subsequent flares are related to each other, and we can identify evidence of flare clustering. Moreover, by comparing the distribution with extreme cases (where all the flares are exactly correlated or none of the flares are correlated) we can determine the fraction of clustered flare events. In addition by varying the lookahead with higher dimensions, we can get some estimate of the size of the clusters. In all cases, features of the distribution functions are compared with surrogate data sets constructed using Bayesian block (BB) decomposition, and the significance of features is based on differences from ensemble averages obtained from multiple realizations of the surrogates.

2. Data Set and Methodology

The flare waiting time data used in our analysis are the same as that used in Snelling et al. (2020). We obtained the solar flare data from the geostationary operational environmental satellite (GOES) catalog of flares from 1975 to 2017, available from https://www.ngdc.noaa.gov/stp/solar/solarflares.html. We considered only flares of class C or higher with a peak flux greater than 1.4×10^{-6} W m^{-2} because of the difficulty of detecting flares with fluxes below class C (Snelling et al. 2020). Event times were taken to be the time of maximum flux during bursts that fit the above criteria. From the sequence of flaring event times (t_1, t_2,…,t_j,…,$t_{N_{\rm flare}}$), we constructed a sequence of 71,587 waiting times (Δ_1, Δ_2,...,Δ_j,...,Δ_{N-1}), where $\Delta_j = t_{j+1} - t_j$. That is, the waiting time is defined as the time interval between two successive events.

The data are analyzed by constructing cumulative distribution functions (CDFs) of flares' waiting times. The CDFs are constructed using standard n-dimensional histogram algorithms developed for MATLAB (histcnd). Given the somewhat exponential distribution of data points, the data is binned in logarithmic bins to maintain a roughly equivalent number of data points in each bin. Different methods have been proposed to estimate the bin size, such as the use of Doane's rule (Doane 1976) based on the Sturges method.

Ideally, there should generally be at least five data points per bin to maintain reasonable statistics. While empty bins do not affect discriminating statistics such as mutual information, too many singly occupied bins lead to inaccuracies in the mutual information. Therefore, bins are selected to be large enough to minimize the number of singly occupied bins, and bin selection is optimized as discussed in Snelling et al. (2020).

The CDF obtained from the data is obtained for the occurrence of multiple flares. The CDF of one variable measures the likelihood that a random point in the data set is less than the value of the variable. For example, the highest value in the data set will have a CDF of one (all points will be less than the highest value) and the lowest value will have a CDF of zero (no points in the data set will be less than the lowest value). Given two values from two data sets, their CDF gives the probability that any two given points from the data sets will be less than the respective given values. For example, if we consider two random variables X and Y where X is the waiting time of the $n^{\rm th}$ flare in the sequence and Y is the waiting time of the $n + p^{\rm th}$ flare in the sequence then we obtain CDF(X, Y), which will measure the fraction of flares with $\Delta_n < X$ and $\Delta_{n+p} < Y$. We also obtain CDF(X, Y, Z), which measures the fraction of flares with $\Delta_n < X$, $\Delta_{n+p} < Y$, and $\Delta_{n+q} < Z$. Clustering of flares can be recognized when there is an elevation in the number of flares compared with the situation where the flare sequence is drawn from a nonstationary Poisson process.

For comparison with the CDF of the data, we construct ensembles of surrogates that satisfy the null hypothesis of a nonstationary Poisson process. The CDF of the original data is compared with the CDF of the surrogate data by taking the difference. This difference indicates where the distribution of waiting times has a higher probability of occurring as a grouped sequence than would be expected if the data were random. In essence, when this difference is large and positive, it means that there is a cluster of flares that occur in a well-ordered sequence. Negative values correspond to a reduced probability of flares occurring together compared with a nonstationary Poisson process. To check that the results are meaningful, we also provide the significance S obtained from

$$S = \frac{|\mathrm{CDF(data)} - \langle \mathrm{CDF(surrogates)} \rangle|}{\sigma_{\mathrm{CDF(surrogates)}}}.$$

3. Creating Nonstationary Poisson Surrogates Using Bayesian Block Decomposition

The surrogate data may be constructed in multiple ways giving similar results. Moon et al. (2001) created surrogates based on a time variable rate obtained by averaging the flare rate over a sliding time window of a few days. The BB algorithm (Wheatland 2000a) is a nonparametric method that can also be used to find an optimal binning for a set of values without imposing a uniform bin width. A nonstationary Poisson process may be subdivided into time intervals where the observed event occurrence is consistent with a (constant rate) Poisson process. This time-dependent process consists of piecewise stationary processes and can be characterized with Bayesian statistics. Thus, these time intervals are characterized by a rate (stationary) and a duration and are called Bayesian blocks. This paper uses the same method developed by Scargle (1998) and used in Snelling et al. (2020) to perform a BB decomposition. Surrogate waiting time sequences are obtained by randomizing the data within each

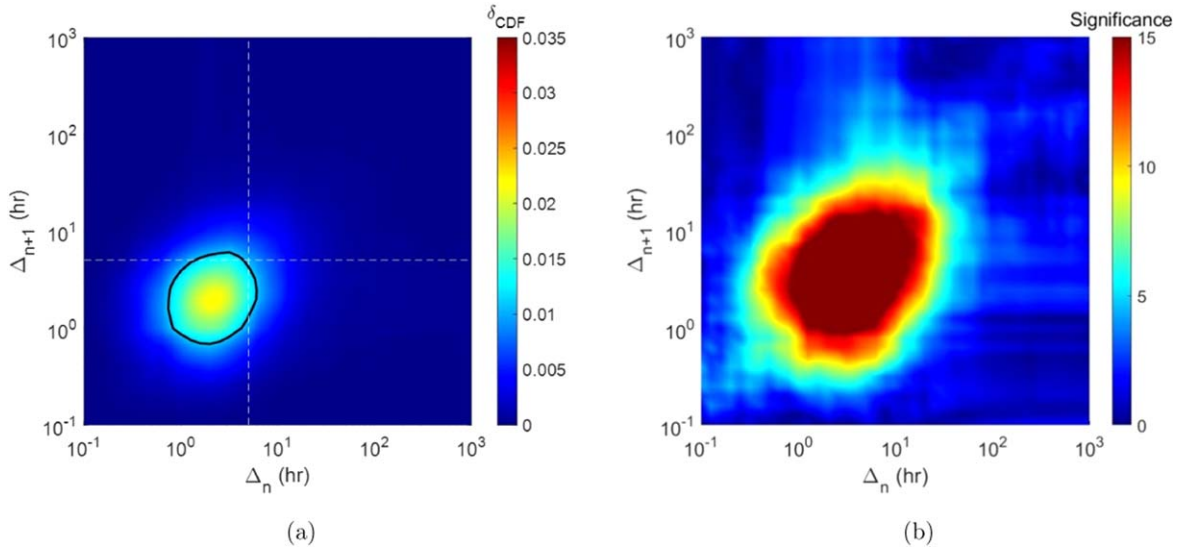

Figure 1. (a) Difference in the joint probability of two waiting times $\delta_{CDF} = CDF(data) - \langle CDF(surrogates) \rangle$ when comparing flares with lookahead $p = 1$ ($\langle A \rangle$ is an ensemble average of property A over the surrogates). The black contour shows $\delta_{CDF} = 0.5 \max(\delta_{CDF})$, and the white dashed lines show the mean waiting time. (b) Statistical significance of δ_{CDF}.

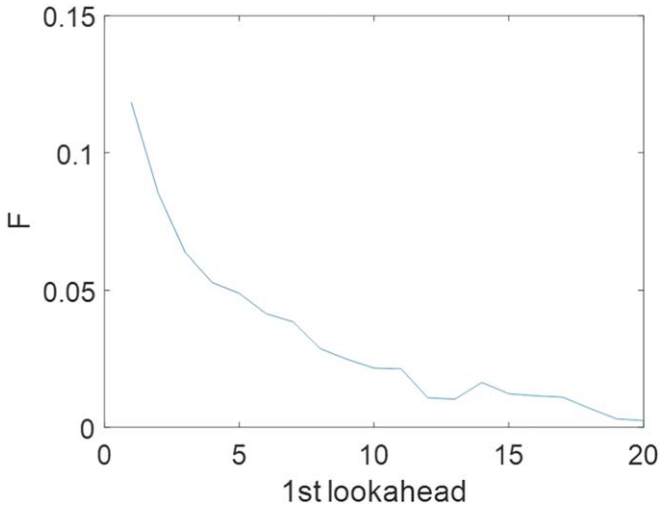

Figure 2. Fraction $F(p) = \max(\delta_{CDF}(p)) / \max(\delta_{CDF}(p = 0))$ as a function of lookahead p (when $p = 0$ all data are completely correlated).

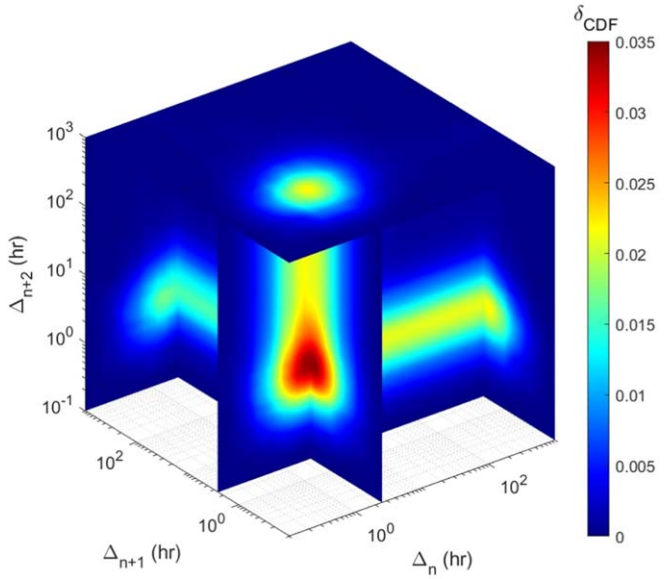

Figure 3. δ_{CDF} of three sequential waiting times. The data show an enhancement in the number of triplet events with similar waiting times around 3 hr.

Bayesian block. When considering the joint CDFs, the number of sequences required for good statistics generally depend on the amount of data available for each histogram cell of the CDF. In general, where differences are large there are plenty of data and 50 surrogates are adequate, while for sparsely occupied cells 1000 surrogates are adequate. It is to be noted that the sparsely occupied cells at long waiting times all have CDFs very close to 1, so differences between the distributions are negligible where the significance is low. In our analysis, we generally use 50 surrogates when comparing CDFs, but for verifying significance we use 1000 surrogates.

4. Results and Discussions

We first analyze the difference in the joint probability of two waiting times $\delta_{CDF} = CDF(data) - \langle CDF(surrogates) \rangle$, where $\langle A \rangle$ is an ensemble average of property A over the surrogates.

The results of this analysis are shown in Figure 1 when comparing flares with lookahead $p = 1$. Panel (a) shows δ_{CDF} and panel (b) shows the statistical significance of δ_{CDF}. It is apparent from panel (a) that there is a higher probability of two subsequent flares each having a waiting time of around 3 hr than would be expected if the flares occurred randomly (considering a slowly changing flare rate). The black contour shows where the probability drops to half the maximum value; $\delta_{CDF} = 0.5 \max(\delta_{CDF})$. This elevated probability is clear evidence that flares are clustered with a short-term memory. For our purposes, we will describe a cluster as a group of flare events in the sequence having a similar waiting time. For comparison, the mean waiting time of the data set (4.8 hr) is plotted as a dashed white line showing that the clustered flares

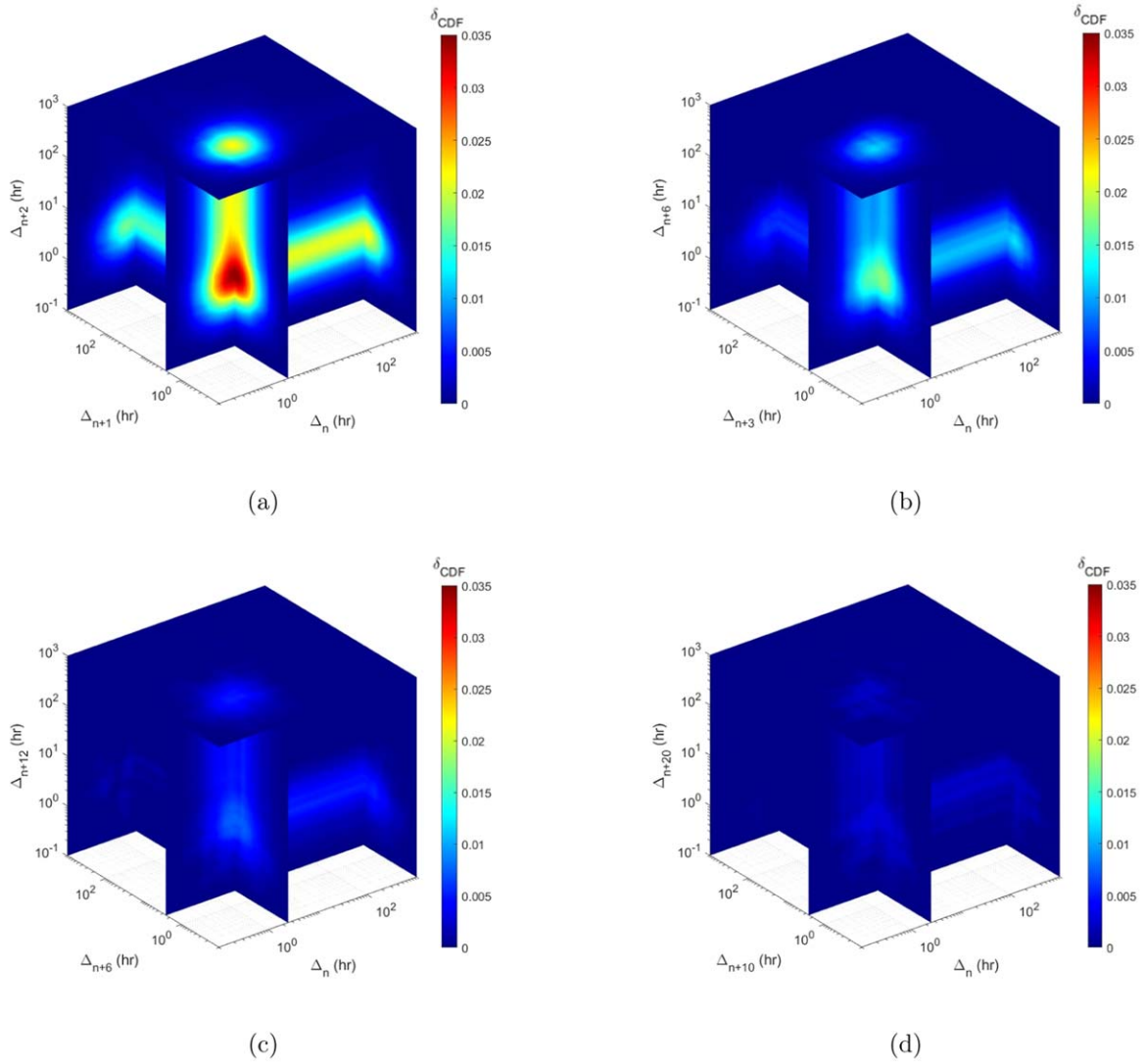

Figure 4. Changes in $\delta_{\rm CDF}$ with lookahead p equal (a) 1, (b) 3, (c) 6, and (d) 10 for the flare sequence of three waiting times.

occur with a higher frequency than the mean flaring rate. That is, the clusters identified are quasi-periodic bursts of flares.

The significance shown in panel (b) shows that the peak of panel (a) has a high significance (clearly differentiated from the surrogate data). The only regions of the plot where the significance is small is in regions where we cannot tell any difference between the actual data and the surrogate data. Therefore, we can generally say that all the features seen in panel (a) are a good representation of the difference between the CDF of the actual data and surrogates.

It should be noted that the timescale of the related pairs of flare waiting times in our study differs substantially from the clusters identified by Wheatland et al. (1998), which had timescales ranging from 10 s–10 minutes. Those clusters were identified based on an overabundance of short waiting times compared with a simulated time variable Poisson process. In our analysis, the surrogate data always has an identical distribution of waiting times as the original data, so differences in the distributions can only be related to time ordering of the waiting times. Our analysis shows that while there may be an abundance of short waiting times (less than 10 minutes) those waiting times apparently do not have a stronger tendency to occur in sequence than waiting times drawn from the same Bayesian blocks.

While the significance measures how different CDF(data) is from CDF(surrogates), it is also of interest to determine how the peak CDF(data) differs from the CDF if the data were perfectly correlated. For comparison, if $p = 0$ we obtain the CDF when all data are completely correlated. In Figure 2 we show the fraction $F(p) = \max(\delta_{\rm CDF}(p))/\max(\delta_{\rm CDF}(p = 0))/$ as a function of lookahead p. What this plot shows is that at $p = 1$, about 12% of the waiting times at around 3 hr appear to be related to each other and that fraction drops substantially as a function of lookahead. It should be noted that this behavior is very similar to the behavior of the mutual information as a function of lookahead obtained by Snelling et al. (2020).

In Figure 3 we examine the CDF of three waiting times that occur in sequence. This figure shows three interesting features. First, there is a peak in the distribution around 3 hr suggesting that, in fact, not only is there is a higher probability of two sequential waiting times having values around 3 hr, but actually there is also an elevated probability that there is a sequence of three waiting times having values around 3 hr compared with what would be expected if the waiting times were randomly distributed. This result goes

beyond the analysis of Snelling et al. (2020), which only considered the relationship between pairs of flares and not triads. A second feature of the distribution are columns of probability emanating from the central peak. These columns represent data where two of the random variables are related, but the third variable is independent. So, the column that runs vertically along the Δ_{n+2} axis represents data where Δ_n and Δ_{n+1} are related, but independent of Δ_{n+2}. In other words, those represent clusters of two similar flare waiting times close to 3 hr followed by a random waiting time. Not surprisingly a similar column is seen along the Δ_n axis also representing a random flare followed by two flares having waiting times close to 3 hr. So the columns represent clusters of length two and the central peak represents clusters of at least length three. Finally, the third column along Δ_{n+1} appears weaker than the other two columns. That column corresponds to a flare having a waiting time about 3 hr followed by a random waiting time followed by another flare having a waiting time around 3 hr. The difference between this column and the other two columns is further evidence of clustering because the two stronger columns detect the edges of the clusters (contributing for every cluster) while the weaker column only detects the density of clusters.

We can get some idea about the extent of the clusters by looking at how the three point CDF varies with lookahead p. Clusters will be characterized by edge events (where the cluster starts and ends) and central events that occur between the edge events. For example if there are clusters of length five, we would expect groups of five events to have similar waiting times. As an example, the sequence of events $\{\Delta_5, \Delta_6, \Delta_7, \Delta_8, \Delta_9\}$ may have similar waiting times. Here we refer to Δ_5 and Δ_9 as edge events and Δ_6, Δ_7, and Δ_8 as central events. To explore the size of the clusters, we can look at how CDF(Δ_n, Δ_{n+p}, Δ_{n+2p}) changes as p increases. When there are clusters of length five as in the example, we would expect that for $p = 1$ we would have a contribution from $\{\Delta_5, \Delta_6, \Delta_7\}$, $\{\Delta_6, \Delta_7, \Delta_8\}$, $\{\Delta_7, \Delta_8, \Delta_9\}$ to a strong peak in $\delta_{\rm CDF}$. For $p = 2$ we would expect that $\{\Delta_5, \Delta_7, \Delta_9\}$ still contribute to a central peak. For $p = 3$ we would not expect to see a central peak, but we would expect to see the two columns from the edges $\{\Delta_5, \Delta_8\}$ and $\{\Delta_6, \Delta_9\}$. Figure 4 shows how $\delta_{\rm CDF}$ changes with p for the flare sequence. A strong central peak is seen for $p = 1$ and a reasonably elevated peak is seen at $p = 3$ suggestive that many events are related having a lookahead up to three.

The central peak is lost around $p = 6$ and only the two edge columns are seen, suggestive that only the edges of the clusters can be detected at this lookahead as in the example above. We can conclude that clusters tend to occur in groups of length six. Figure 5 compares the maximum $\delta_{\rm CDF}$ as a function of p confirming a substantial reduction up to $p = 6$.

5. Summary

Overall our analysis supports the idea that flares have a short-term memory, which couples to some extent, the dynamics of groups mostly of lengths ranging up to six subsequent flares. By looking at the differences in the joint CDF of two and three flares we were able to identify how these flares are related and understand better why the mutual information between successive flares was elevated (Snelling et al. 2020). We determined that the occurrence of sequences of two or three flares having waiting times in the range of 2–6 hr is elevated well beyond what would be expected if the flare waiting times were drawn from a nonstationary Poisson process. While the fraction of the flare data participating in the short-term memory is small, it is not negligible and corresponds to about 12% of flares at these timescales. From the joint CDF of three flares, we also were able to determine that the extent of the clusters of related flares are typically found in groups of six flares. These results are consistent with prior work suggesting short-term memory of solar flares (Snelling et al. 2020; Aschwanden & Johnson 2021). These findings could be further explored in the context of other statistical methods (Feigelson et al. 2022) and may be useful in developing predictive models for flare events.

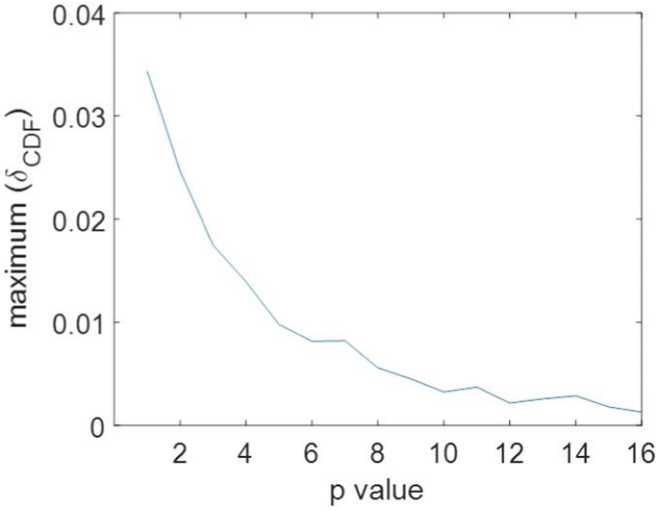

Figure 5. Maximum $\delta_{\rm CDF}$ of three waiting times that occur in sequence as a function of lookahead p.

Three-dimensional MHD simulations of the upper convection zone and photosphere/chromosphere modeling flux emergence suggest that current layers develop and ejection of plasmoids out of the current layers leads to patchy reconnection and the spontaneous formation of clusters of microflares (Archontis & Hansteen 2014). Although the flares identified in our analysis are generally much stronger (C class and larger), the general notion that recurrent quasi-periodic flare emissions can result from dynamically driven interconnected networks of magnetic flux that could lead to flare sympathy (Wheatland & Craig 2006), should be further explored.

This work is supported by NASA grants NNX16AQ87G, 80NSSC21K1678, 80NSSC20K0355, 80NSSC20K0704, 80NSSC22K0515, and NSF AGS grant 2131013.

ORCID iDs

Elmer C. Rivera https://orcid.org/0000-0001-7747-9804
Jay R. Johnson https://orcid.org/0000-0002-9562-1103
Simon Wing https://orcid.org/0000-0001-9342-1813

References

Archontis, V., & Hansteen, V. 2014, ApJL, 788, L2
Aschwanden, M. J. 2019, ApJ, 887, 57
Aschwanden, M. J., & Johnson, J. R. 2021, ApJ, 921, 82
Aschwanden, M. J., Johnson, J. R., & Nurhan, Y. I. 2021, ApJ, 921, 166
Aschwanden, M. J., & McTiernan, J. M. 2010, ApJ, 717, 683
Charbonneau, P. 2020, LRSP, 17, 4
Doane, D. P. 1976, Am. Stat., 30, 181

Feigelson, E. D., Kashyap, V. L., & Siemiginowska, A. 2022, arXiv:2203.08996
Gilroy, K. L., & McCuen, R. H. 2012, JHyd, 414-415, 40
Hong, L.-L., & Guo, S.-W. 1995, BuSSA, 85, 814
Johnson, J. R., & Wing, S. 2005, JGRA, 121, 9378
Johnson, J. R., & Wing, S. 2014, GeoRL, 41, 5748
Johnson, J. R., Wing, S., & Camporeale, E. 2018, AnGeo, 36, 945
Lei, W., Li, C., Chen, F., et al. 2020, MNRAS, 494, 975
Lepreti, F., Carbone, V., & Veltri, P. 2001, ApJL, 555, L133
Li, C., Zhong, S., Xu, Z., et al. 2018, MNRAS, 479, L139
Moon, Y.-J., Choe, G., Yun, H., & Park, Y. 2001, JGRA, 106, 29951
Nurhan, Y. I., Johnson, J. R., Homan, J. R., Wing, S., & Aschwanden, M. J. 2021, GeoRL, 48, e94348
Scargle, J. D. 1998, ApJ, 504, 405
Snelling, J. M., Johnson, J. R., Willard, J., et al. 2020, ApJ, 899, 148
Toriumi, S., & Wang, H. 2019, LRSP, 16, 1
Wheatland, M. 2000a, SoPh, 191, 381
Wheatland, M., & Litvinenko, Y. E. 2002, SoPh, 211, 255
Wheatland, M. S. 2000b, ApJL, 536, L109
Wheatland, M. S., & Craig, I. J. D. 2006, SoPh, 238, 73
Wheatland, M. S., Sturrock, P. A., & McTiernan, J. M. 1998, ApJ, 509, 448
Wing, S., Brandt, P., Mitchell, D., et al. 2020, ApJ, 159, 249
Wing, S., & Johnson, J. R. 2019, Entrp, 21, 140
Wing, S., Johnson, J. R., Camporeale, E., & Reeves, G. D. 2016, JGRA, 121, 9378
Wing, S., Johnson, J. R., Turner, D. L., Ukhorskiy, A. Y., & Boyd, A. J. 2022, JGRA, 127, e30246
Wing, S., Johnson, J. R., & Vourlidas, A. 2018, ApJ, 854, 85