The Astrophysical Journal Supplement Series, 265:28 (16pp), 2023 March https://doi.org/10.3847/1538-4365/acb43a

A Systematic Magnetic Polarity Inversion Line Data Set from SDO/HMI Magnetograms

Anli Ji[1], Xumin Cai[1], Nigar Khasayeva[1], Manolis K. Georgoulis[2], Petrus C. Martens[3], Rafal A. Angryk[1], and Berkay Aydin[1]
[1] Department of Computer Science, Georgia State University, Atlanta, GA, USA; aji1@student.gsu.edu, xcai3@student.gsu.edu, nkhasayeva1@student.gsu.edu, angryk@cs.gsu.edu, baydin2@gsu.edu
[2] Research Center for Astronomy and Applied Mathematics, Academy of Athens, Greece; manolis.georgoulis@academyofathens.gr
[3] Department of Physics & Astronomy, Georgia State University, Atlanta, GA, USA; martens@astro.gsu.edu
Received 2022 June 2; revised 2022 December 13; accepted 2023 January 4; published 2023 March 8

Abstract

Magnetic polarity inversion lines (PILs) detected in solar active regions have long been recognized as arguably the most essential feature for triggering instabilities such as flares and eruptive events (i.e., eruptive flares and coronal mass ejections). In recent years, efforts have been focused on using features engineered from PILs for solar eruption prediction. However, PIL rasters and metadata are often generated as by-products and are not accessible for public use, which limits their utilization in data-intensive space weather analytics applications. We introduce a large-scale publicly available PIL data set covering practically the entire solar cycle 24 for applying to various space weather forecasting and analytics tasks. The data set is created using both radial magnetic field (B_r) and line-of-sight (B_LoS) magnetograms from the Solar Dynamics Observatory's Helioseismic and Magnetic Imager Active Region Patches (HARP) that involve 4090 HARP series ranging from 2010 May to 2019 March. This data set includes three PIL-related binary masks of rasters: the actual PILs as per the spatial analysis of the magnetograms, the region of polarity inversion, and the convex hull of PILs, along with time-series-structured metadata extracted from these masks. We also provide a preliminary exploratory analysis of selected features aiming to correlate time series of feature metadata and eruptive activity originating from active regions. We envision that this comprehensive PIL data set will complement existing data sets used for space weather forecasting and benefit research in related areas, specifically in better understanding the PIL structure, evolution, and role in eruptions.

Unified Astronomy Thesaurus concepts: Astronomy software (1855)

1. Introduction

A polarity inversion line (PIL) in solar active regions, almost always lying underneath solar filaments and prominences, can be thought of as a line (or curve, in general) separating positive and negative polarities of magnetic flux of the Sun. Characterized by both velocity and magnetic shear, strong PILs are literally a shear layer in nature, showcasing the layer where the magnetic polarity reverses. For more than half a century, a strong correlation between intense PILs and eruptive solar activity has been inferred (e.g., Severny 1964; Heyvaerts et al. 1977; Hagyard et al. 1984; Zirin & Liggett 1987; van Ballegooijen & Martens 1989; Antiochos et al. 1999; van Driel-Gesztelyi & Green 2015; Georgoulis et al. 2019; Toriumi & Wang 2019, and many others). In more recent decades, this collective knowledge has transpired into efforts using PIL features, often automatically identified (e.g., Jones 2004; Padinhatteeri et al. 2015) for solar flare and eruption forecasting, as it has been conceived that such predictions can benefit from a number of primary PIL features. Although different methods for detecting PILs from full-disk solar magnetograms or active region cutouts exist (e.g., Mason & Hoeksema 2010; Falconer et al. 2014; Sadykov & Kosovichev 2017; Park et al. 2018; Korsós et al. 2019; Wang et al. 2020; Cicogna et al. 2021; Sun et al. 2021), most of these data sets are not publicly available for widespread use in data-intensive, machine-learning, or deep learning–based studies.

In a previous work (Cai et al. 2020), we introduced a detection framework that identifies PIL masks from the Solar Dynamics Observatory's Helioseismic and Magnetic Imager (SDO/HMI) active region magnetogram cutouts. The primary detection approach is based on a strength-filtered edge detection technique, together with a masked total unsigned flux criterion coupled with PIL size filter threshold. Similar to the techniques used in Volobuev (2011) and Wang et al. (2019), we use the magnetic field strength threshold to identify the positive and negative polarity regions and use the size filter to remove less critical (i.e., weaker) portions of the lines to preserve the topology and original shapes.

In this work, we introduce a novel PIL detection data set from SDO/HMI active region cutouts that can be used for a wide range of space weather forecasting problems. The data set (available at doi:10.7910/DVN/BKP1RH) involves 4090 definitive active region HMI Active Region Patches (HARP) series (for complete information on SDO/HMI HARP, see Hoeksema et al. 2014). Such series are made available at a cadence of 720 s and are mapped with the Lambert cylindrical equal area (CEA) projection (Snyder 1987), spanning between 2010 May and 2019 March. For each HARP series, we generate relevant physical and shape-based metadata feature time series (i.e., PIL size, the region of polarity inversion [RoPI], the masked unsigned flux of enclosing PIL, convexity, eigenvalues, fractal dimension, and Hu moments). These are accompanied by a detection quality flag indicating a successful detection, as per our criteria and settings. In addition, for each

active region patch in the series, we generate three binary masks in image format with the same pixel size as the corresponding magnetogram so that they can be readily mapped to the original images.

The rest of the paper is organized as follows. In Section 2, we introduce the data set for the PIL of the SDO/HMI magnetogram. In Section 3, we give an example of an exploratory analysis on the PIL data set, in both abstract and detailed terms. In Section 4, we provide our concluding remarks and an outlook for future work.

2. PIL Data Set

2.1. Data Collection and Preprocessing

We applied our detection method in Cai et al. (2020) to both the radial magnetic field (*B*_r) and line-of-sight (*B*_LoS) magnetogram components obtained from the HARP data product (Hoeksema et al. 2014) stored in FITS format. We primarily use the definitive series with a 720 s cadence and map it to the Lambert CEA projection, where each CEA pixel roughly covers the same physical area on the Sun. Considering the magnetic field projection effects, we carry out validity testing for magnetogram patches and apply the detection method only to the qualified HMI magnetogram patch whose central angles (in heliocentric coordinates) of the bounding box corners are less than 70° east–west.

2.2. Detection Method

Drawing from our previous detection method in Cai et al. (2020), we start by identifying the positive (i.e., upward-pointing) and negative (i.e., downward-pointing) polarity regions by generating and applying two binary masks to each of the opposite polarities with a magnetic field strength threshold of ±100 G. Then, for all detected polarity regions, we further apply a preliminary polarity region size filter of 14 pixels (i.e., ∼1.86 Mm^2 of the photospheric area) to reduce the noise-like smaller regions. Next we obtain the boundaries of positive and negative polarity regions separately by using the Canny edge operator (Canny 1986), which employs a Gaussian filter to smooth the magnetogram images and to detect the overall contours of the polarity regions. The RoPI is then defined as the intersection region between spatially buffered edges where positive and negative polarity regions are adjacent. Here we spatially buffer the edges of negative and positive polarity regions in binary images with a size 4 kernel to reasonably represent the area around a given PIL. To generate the candidate PILs, we utilize an unsigned magnetic field strength filter on the areas covered by disconnected RoPIs, which ensures that the selected candidate PILs represent at least 95% of the unsigned magnetic flux contained in RoPIs and removes the candidate PILs with a lower flux buildup around them. Following Serra (1983), we then apply a morphological thinning operation to the candidate RoPIs to eliminate the majority of the foreground pixels and to preserve the topology of the primary shapes of PILs. Finally, we apply a size filter of 14 CEA pixels (∼5 Mm in photosphere) to remove relatively small inversion lines and retain those cleaner and more significant ones. This choice of size threshold ensures that any structure that may be affected by the quiet Sun, featuring granules and complexes, is not retained.

2.3. PIL Data Product

2.3.1. Data Formats and Availability

For each active region magnetogram, the PIL data product we create comprises two parts: binary masks and metadata. For each magnetogram, our detector produces three PIL-related binary masks: the thinned PIL (i.e., the actual PILs as per the spatial analysis of the magnetograms), the RoPI, and the convex hull of PIL (i.e., the minimal area covering all PILs, indicating convex closure of the set of detected PILs). The binary masks are stored in PNG format within the same size element as the pixel size of the corresponding magnetogram. Thus, the binary mask can be readily mapped onto the original HMI magnetogram raster, avoiding any image transformation procedure. Each PIL-related binary mask is uniquely labeled by a combination of

1. A unique active region identifier (HARP number—HARPNUM).
2. A specific timestamp corresponding to the original magnetogram.
3. The type of the binary mask (i.e., thinned PIL, RoPI, or convex hull of polarity inversion).
4. Detection parameters that include initial magnetic field strength threshold (in Gauss) and a PIL size filter threshold (in pixels).

We present an illustrative example of generated binary masks in Figure 1, overlaid on the original magnetogram patches. The created binary masks provide researchers and practitioners with a useful resource for exploring the structure and evolution of PILs. The users of the data set can remap the PIL masks to the corresponding magnetogram for visual analysis or extract the application-dependent PIL features to utilize as inputs for a variety of learning tasks or to help address a range of physical questions.

In addition to the rasters of PIL-related binary masks, we generate some well-known physical-based shape descriptors. This information is stored in CSV text format and can be viewed as timestamped time series data based on the original HARP series of a given active region. In Table 1, we list the PIL-related features and shape descriptors provided in the data set.

In particular, the number of PILs represents the count of disconnected PILs in the magnetogram patch. The PIL size represents the sum of all pixel counts contained in the PIL mask, where each CEA pixel has a characteristic length of ∼362.5 km and covers approximately 1.33×10^5 km^2 (Bobra et al. 2014). For the RoPI, we count the number of CEA pixels covered by the RoPI. The masked unsigned flux of PIL is calculated by the sum of $|B_{los}|$ contained in the RoPI multiplying by the pixel area, which provides magnetic flux in physical units. The fractal dimension (Mandelbrot 1985) evaluates the self-similarity in the PIL mask, and we calculate it using the box-counting method provided in Mandelbrot (1985) for each PIL. We also provide eigenvalues (Woodcock 1977) of the PIL covariance matrix (i.e., covariance of individual pixels), which can help measure the ratio of principal axes lengths between PILs. The convexity (Mora & Kwan 2000) refers to the ratio of the perimeter of the PIL convex hull area to the sum of pixels in the PILs (i.e., PIL size). Hu moments (Ming-Kuei 1962), another set of well-known shape descriptors, comprise seven moment values that are translation, scale, and rotation invariant.

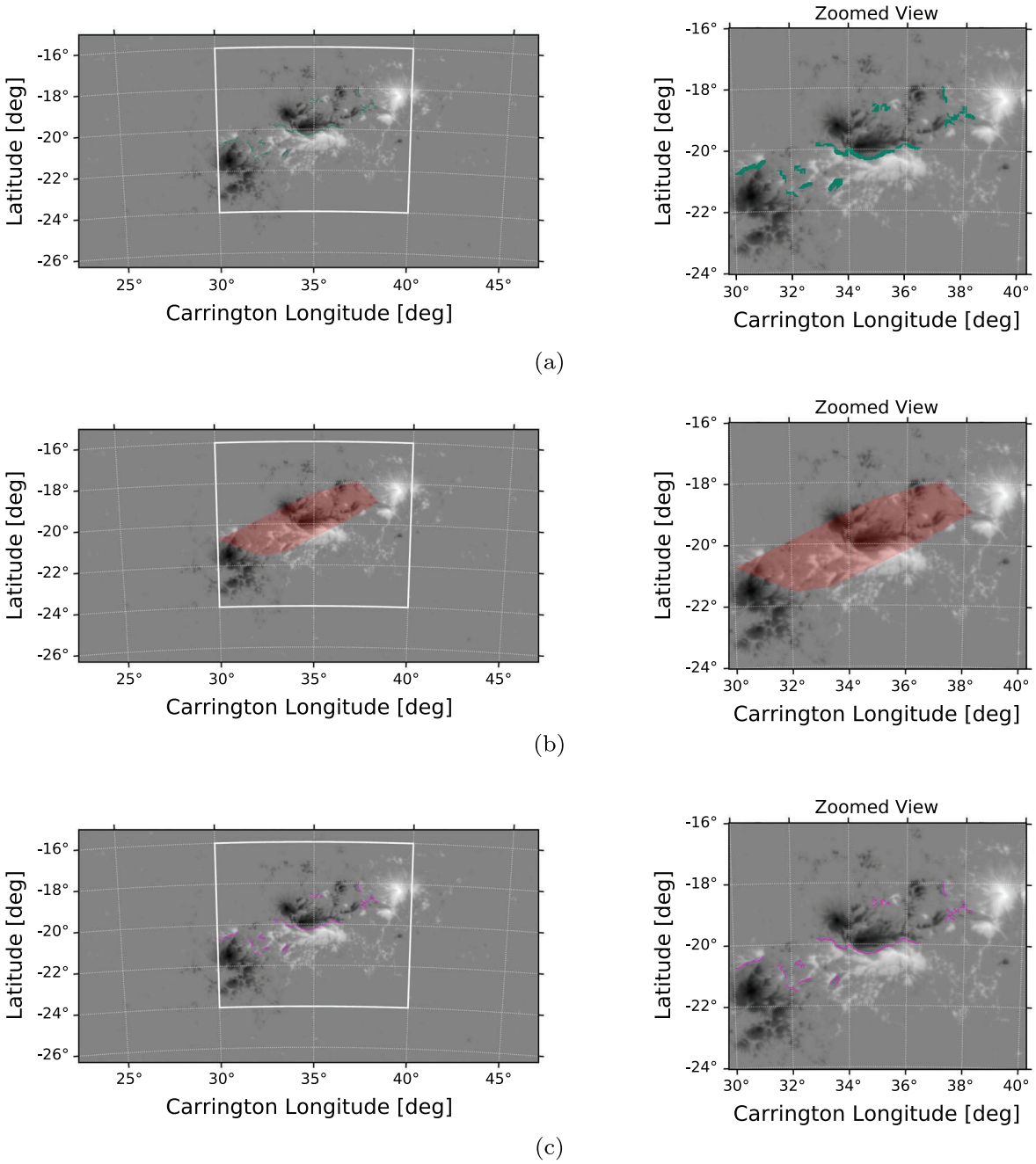

Figure 1. Binary masks overlaying the HMI magnetogram patch of NOAA Active Region 11158 (HARP 377) on 2011 February 14 at 04:24:00 UT. The left three figures show the original magnetogram patches with the zoomed-in area, while the right three show the zoomed-in detailed patches. In panel (a), the green area indicates the RoPI binary mask. In panel (b), the red area shows the binary mask of the convex hull of the PIL network in the AR. In panel (c), the pink line represents the detected PIL binary mask.

Additionally, we include a validity flag indicating the quality of generated PILs at each timestamp. There are three different values for the validity flag: data records marked as "D" refer to PIL data successfully detected under the current parameter settings. Data records marked as "WF" indicate that no PILs were detected due to weak flux. Data records marked as "SPE" indicate no detection available due to proximity to the solar limb, that is, the central angle (in heliocentric coordinates) of the bounding box corners is less than 70° east–west. We only produce PIL binary masks and metadata time series thereof for PILs marked by a "D" validity flag.

3. Synoptic Statistical Analyses of Polarity Inversion Lines

We conduct three exploratory analyses using features of our PIL data set derived from *B*_r magnetograms to reveal and better understand the relationship between PILs and eruptive activity in solar active regions. A respective preliminary study using features from our PIL data set generated by HMI *B*_LoS magnetograms is presented in the Appendix A.

In the first part of our analysis, we explore the relationship between flare and eruption generation abilities of individual active regions with respect to the evolution of PIL-related time series metadata. For the second analysis,

Table 1
Total of 14 PIL Features Provided in CSV File Metadata in the Form of Time Series

Features[a]	Description
Number of PILs (1)	Number of disconnected PILs in the PIL binary mask
PIL size (1)	Total number of CEA pixels contained in the PIL binary mask
RoPI Area (1)	Total number of CEA pixels contained in the RoPI binary mask
Masked Unsigned Flux of PIL (1)	Sum of $\lvert B_{\rm los}\rvert$ enclosed by RoPIs (gauss)
Fractal Dimension (1)	Shape descriptor evaluating the self-similarity in the PIL
Eigenvalues (1)	λ_1 and λ_2, used to measure the ratio between the length of principal axes
Convexity (1)	Ratio of the perimeter of the PIL convex hull to the sum of CEA pixels of PILs
Hu Moments (7)	Seven image moments from Ming-Kuei (1962) (i.e., translation, scale, and rotation invariant)

Note.
[a] The numbers in parentheses represent the amount of subfeatures included in each CSV file. For example, there are seven subfeatures in total for Hu moments.

we focus on discovering the evolution of PILs prior to the occurrence of a flare event in terms of different observational window settings. In the last analysis, we aim to understand the PIL characteristics as well as their discriminative correlations among different flare intensities.

For each active region analyzed, we slice through the span of the time series so that each slice represents an observational window containing 24 hr worth of data prior to each flare event. We disqualify and reject instances with less than 24 hr of observations as well as those overlapping with Solar and Heliospheric Observatory/Large Angle and Spectromeric Coronagraph Experiment downtime. Moreover, the labels for each time series slice are further verified with three independent resources that provide coronal mass ejection (CME) to flare association information: the Space Weather Database of Notifications, Knowledge, Information (DONKI; from 2010 to 2021),[4] Coordinated Data Analysis Workshops (from 1997 to 2018; Gopalswamy et al. 2009), and the Low Coronal Event Catalog (from 2007 to 2017; Murray et al. 2018). We implement a custom verification schema among our existing CME-flare association list (Ji & Aydin 2022) and all the instances from the above three different catalogs. With the corresponding start and peak times, every flare record can have as many as three additional information resources in parallel. A supplementary CME association confidence indicator, ranging from −3 to 3, is computed based on the source availability. The positive confidence score shows the availability of additional sources validated for such flare events in addition to the existing CME list, while the negative score shows the source availability only when such CME events have not been recorded in our existing list. In this way, we are able to discriminate between our listed association with the additional CME information and further complement our existing list. In our studies, we only consider events if data are available in both the internal and external resources with a high confidence score (i.e., vconf $\geqslant|2|$). It is worth mentioning that our approach filters out around 10% of the data instances; however, it leads to a data set with only high confidence in the occurrence of CME events. The eliminated instances are stored in a separate filter file for future references. Again, our purpose here is to demonstrate the applicability of our PIL data set and related metadata for solar flare and eruption prediction, not necessarily as an operational forecasting tool, at least not at this initial stage.

[4] https://kauai.ccmc.gsfc.nasa.gov/DONKI/

3.1. Exploratory Analysis 1: Flaring versus Eruptive Active Regions

As part of the first analysis for exploring flare and eruption generation abilities, we label solar active regions based on two different criteria:

1. Flaring active region (i.e., FL_ARs), namely, those associated with at least one C+ class solar flare.
2. Eruptive active region (i.e., ERP_ARs), namely, those associated with at least one C+ class eruptive flare.

Here we defined an eruptive flare as a subgroup of the flaring instances that are correlated with at least one CME event. As mentioned earlier, we only take high-confidence eruptive flares into consideration and exclude the uncertain ones in our analysis. From a total of 9035 HARP series between 2010 May and 2018 December, we have 661 HARP active regions considered as flaring active regions and 168 HARP active regions considered as eruptive active regions. We compute the mean and standard deviation of three selected features (i.e., number of PILs, PIL size, and masked unsigned flux of PIL) over the entire time period and calculate the number of associated flares for each active region. Notice that we only consider active regions with data lasting more than 24 hr and eliminate active regions that do not have enough data as well as those data points that do not pass the detection quality check (magnetic field strength less than our 95% threshold or influenced by severe projection effects). The distributions of selected features in the *B*_r data set and the number of associated flares per day are shown in Figure 2 using scatter plots where each point represents either a flaring or an eruptive HARP active region. All instances from nonflaring active regions are excluded in this part of the experiment.

From Figure 2, we notice that if the number of associated daily flares exceeds 4, active regions are exclusively eruptive, which means that at least one (not necessarily all) of these flares is associated with a CME. It is worth mentioning that we take the overall number of associated flares into consideration; therefore not all the flares from our eruptive active regions are actually eruptive. As for the statistical means shown, we can see that there are certain thresholds for the distinction of flaring and eruptive active regions. For example, if an average of more than 18 PILs are detected (throughout the entire lifespan), then over 92% of these active regions are identified as eruptive. As the average PILs size exceeds 300 CEA pixels, about 83.8% of the active regions are marked as eruptive ones. On the other hand, flaring active regions appear mainly on the lower range of our selected features. For instance, when the standard

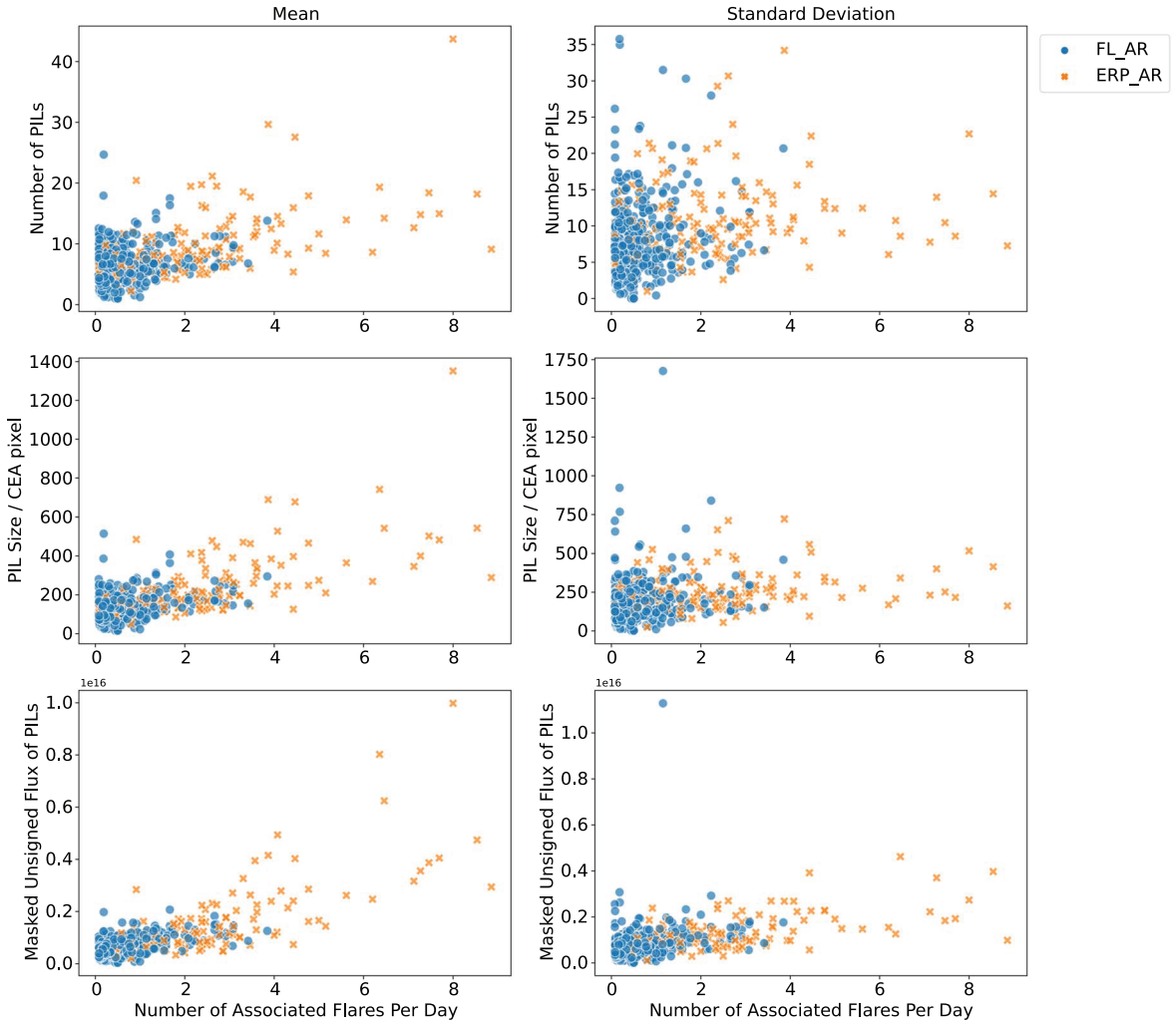

Figure 2. Mean and standard deviation scatter plots of three features (i.e., number of PILs, PIL size, and masked unsigned flux of PIL) derived from the radial magnetic field magnetograms over the entire flaring active region time series in regard to their relationships with the number of associated flares per day.

deviation of the number of PILs drops below 15, around 76% of the marks are considered flaring (but not eruptive) active regions.

3.2. Exploratory Analysis 2: Different Observational Time Period prior to Flare

In the second analysis, we look into the evolution of PILs before the occurrence of a flare and randomly generate an equal amount of instances from nonflaring active regions for maintaining a balanced ratio between flaring and nonflaring instances. Our analysis is conducted in such a way that it includes multiple flares, meaning that more than one flare is included from the same active region. Another analysis using only the largest flare from each active region is presented in Appendix B as well. Besides slicing with only a 24 hr observation window, we also extract the PIL time series data based on a 6 and 12 hr window prior to the event. For each of these time periods, we calculate the statistical measurements (i.e., mean and standard deviation) and present the results using the boxplots shown in Figure 3 for the means and Figure 4 for the standard deviations. By using this graphing technique, we can observe the summary of different flare classes (i.e., minimum, maximum, median, and first/third quartiles) as well as variability outside the upper and lower quartiles (i.e., outliers). Among the different observational windows, while there are major overlaps across different classes of flares, we see a general decreasing trend across M, C, and nonflaring (i.e., NF) classes. This indicates a consistency with the experimental results we described previously: the higher the values of our selected features (i.e., number of PILs, PIL size, and masked unsigned flux of PIL), the more likely the association with larger flares. These simple, preliminary statistical features, even considered alone, could conceivably help predict the occurrence of C+ flares and be useful when predicting the flare intensities.

Although the statistical behavior of the means and standard deviations is similar, a significant difference is that the M class interquartile range of the number of PILs covers a wider range than that of the X class range. This means that the parameter with such characteristics will be less representative compared with the others in terms of distinguishing between M and X class flares. In regard to outliers, there is a slight rise in the standard deviation–based features as we decrease the length of the observational window. This might be due to the facts that

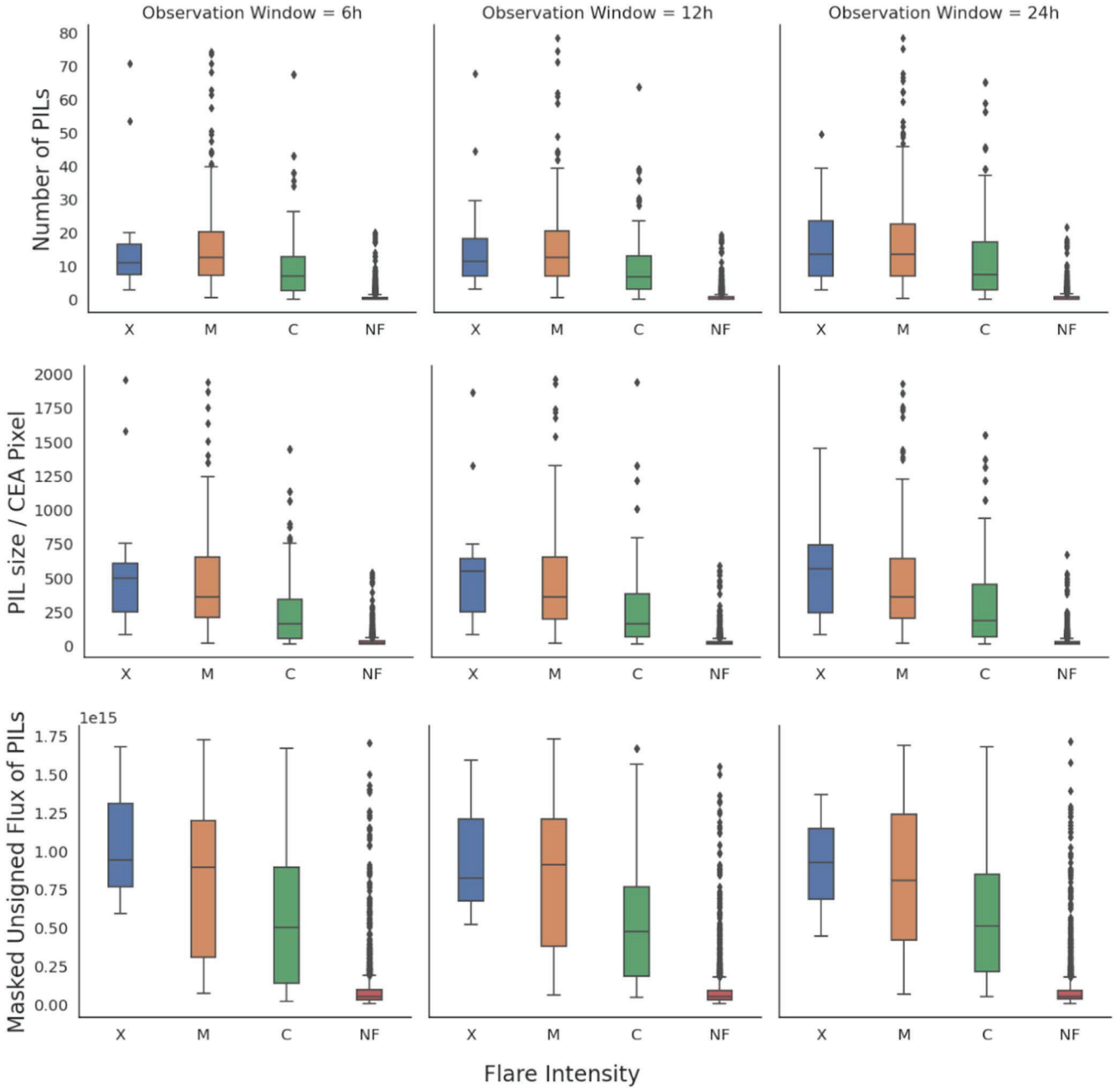

Figure 3. Box plots of means between selected PIL features prior to flaring and nonflaring instances based on Geostationary Operational Environmental Satellite (GOES) flare intensity level for 6, 12, and 24 hr time intervals from the radial magnetic field magnetograms.

outliers are likely to appear on a higher longitude close to the limb area and radial magnetograms tend to be more spatially dispersed with smaller isolated flux regions (where regions contain less emerging accumulated fluxes), causing the number of PILs to fluctuate and resulting in more outliers in the standard deviation.

We believe this overlapping relationship could be caused by the categorical classification where an X1.0 flare is 10 times stronger than an M1.0 flare but probably not distinguishable from an M9.0 flare based on Geostationary Operational Environmental Satellite (GOES) X-ray flux. It is also worth noticing that we do not include relatively smaller flares (i.e., A- and B-class flare) due to the fact that they are severely underdetected. This might create the potential for those high B classes (i.e., B8, B9), including in the nonflaring instances (i.e., NF), to interfere with the C-class statistics.

3.3. Exploratory Analysis 3: Similarity and Discrimination between Flare Intensities

In our third analysis, we explore the characteristics of PILs and how discriminative they are across different flare intensity levels. To pursue this, we have separated flares of GOES classes into five subcategories: C1–C4 for C1.0- through C4. 9-class flares, C5–C9 for C5.0- through C9.9-class flares, M 1–M4 for M1.0- through M4.9-class flares, M5–M9 for M5.0- through M9.9-class flares, and X1+ for flares X1.0 and above. For each category, we examine the distributions of statistical time series features (i.e., mean and standard deviation) for the

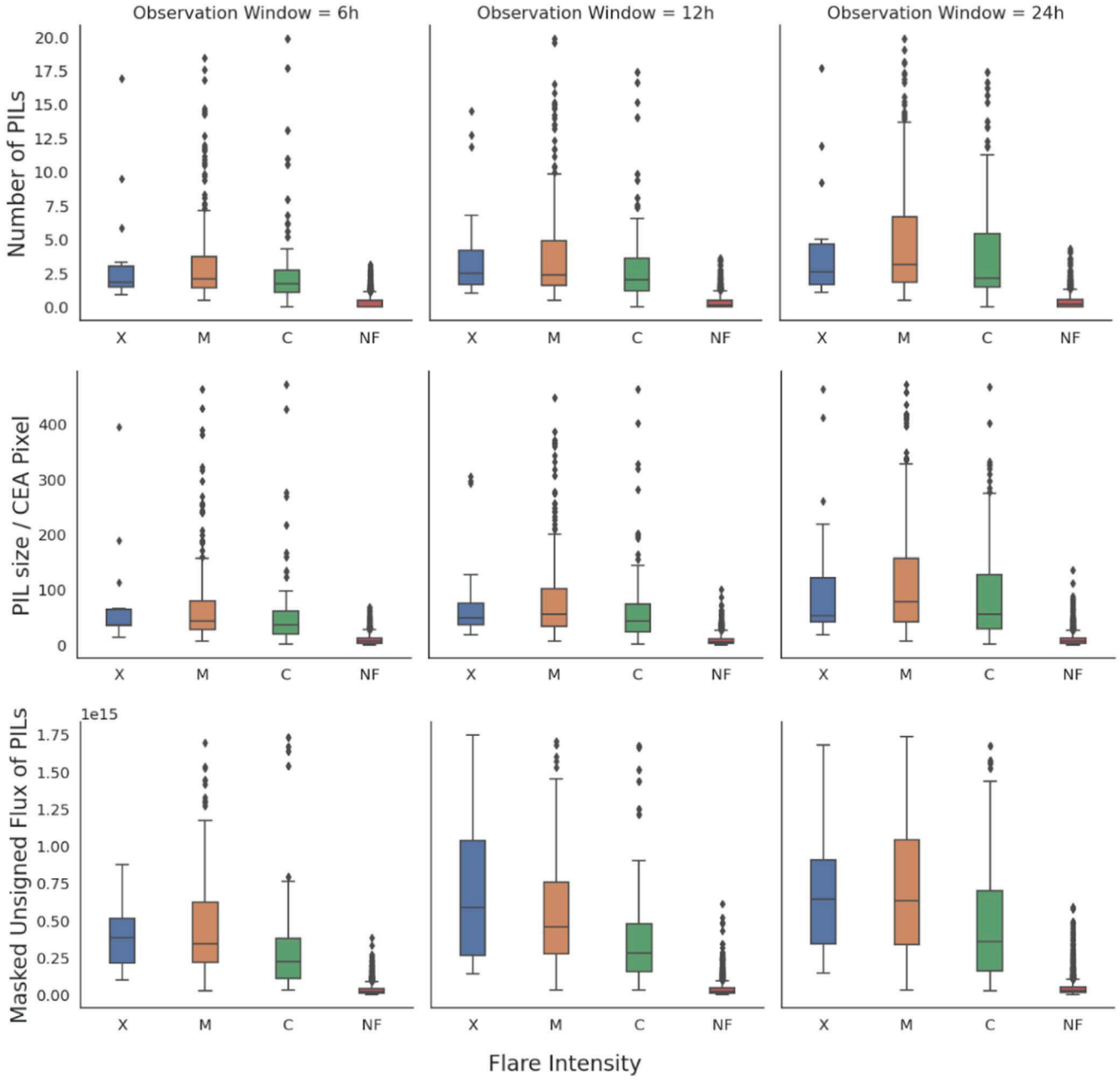

Figure 4. Box plots of standard deviations between selected PIL features prior to flaring and nonflaring instances based on GOES flare intensity level for 6, 12, and 24 hr time intervals from the radial magnetic field magnetograms.

three previously selected PIL features: number of PILs, PIL size, and masked unsigned flux of PILs. We standardize our analysis by using a 12 hr observation window for the preflare time series. To understand the sample distributions of these subcategories, we generate histograms of 20 equal bins for each statistical feature-parameter combination, where the same bin ranges were applied to all subcategories per combination. Such histograms show the sample distributions across, and we evaluate the dissimilarities between these distributions using the discrete version of the Bhattacharyya distance (Kailath 1967) as shown in Equation (1):

$$D_B(H_1, H_2) = -\log \sum_i \sqrt{P_i(H_1), P_i(H_2)}, \quad (1)$$

where $P_i(H_j)$ shows the probability densities for the ith bin of histogram distribution H_j. The distance range is between 0 and 1, meaning that a distance of 0 represents identical distributions, while a distance of 1 represents totally dissimilar distributions. Higher dissimilarity means more separability between the two density distributions. Our results are summarized in Figure 5 as a set of heat maps. The diagonal plots provide the histogram distribution for the five subcategories. Our remarks are then summarized as follows:

1. Histograms in the diagonal often present a normal (i.e., right-skewed) distribution. However, we observe that outliers emerge higher toward values in C5–C9 class histograms (mean series, in particular).

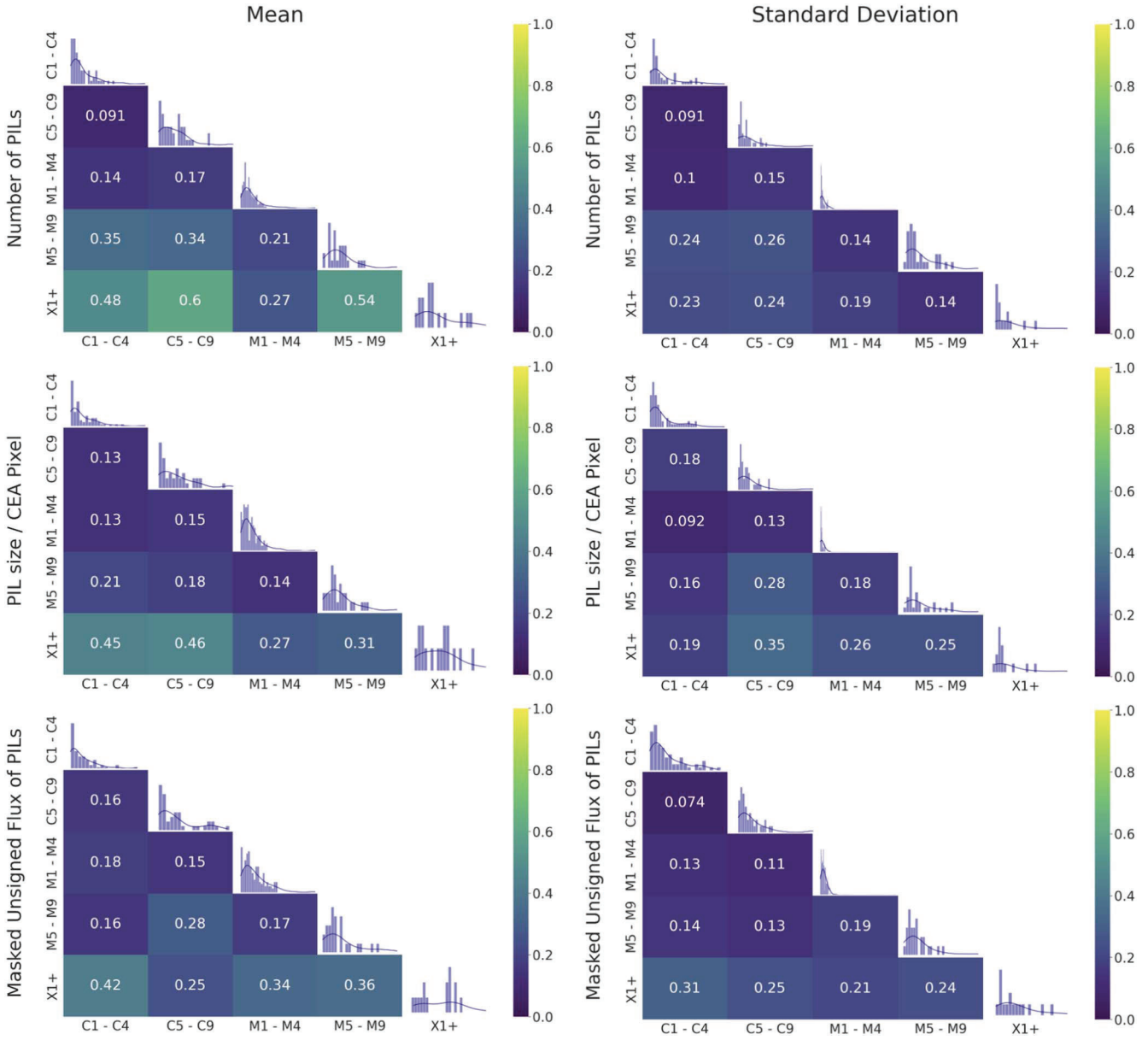

Figure 5. Heat maps for comparing similarity between different flare intensities by using the Bhattacharyya distance from the radial magnetic field magnetograms. The histogram on top of each column shows how flares are distributed within each category.

2. The maximum distance is observed between lowest (C1–C4) and highest (X1+) intensity classes consistently across feature-parameter combinations. The highest distance is achieved for the mean of PIL size and the standard deviation of masked unsigned flux (for C1–C4 and X1+ subcategories).
3. In general, neighboring classes are least separable, as expected. Among these, C-class flares (i.e., C1–C4 and C5–C9 subclasses) and M-class flares (i.e., M1–M4 and M5-M9 subclasses) show the highest histogram similarity.
4. The mean of the time series appears generally more effective than the standard deviation in terms of histogram separability. There are a few exceptions to this: for example, the mean of masked unsigned flux is slightly better than the standard deviation (0.16 versus 0.07) when separating between C1–C4 and C5–C9 in terms of the *B*_r data set.
5. We notice the interesting fact that the histograms of X1+ class flares are more similar to M1–M4 than M5–M9 class flares. We believe that this effect is due to the sizes of the subclass populations and outliers when there are significantly higher numbers of lower-intensity M-class flares (M1–M4), which may occur in swarms prior to, or after, an X-class flare, leading to similarities with a subset of M1–M4 flares.

We underline an established notion within the solar physics community, namely, that the mechanism leading to flares (i.e., magnetic reconnection) is the same regardless of flare size, while the size itself is subject to the availability of different amounts of magnetic energy to fuel the flare in an overall self-similar, stochastic manner (see, e.g., Vlahos & Georgoulis 2004; Campi et al. 2019). The exploratory analysis presented here pertains to a single time series parameter, while complex and more sophisticated predictive analytics should consider

using all available/viable parameters for training flare forecasting models. Overall, we see that even our simple statistical features from preflare time series of PIL-related features can show promise in separating flare intensity subclasses. Promise may be further shown toward a categorical separation of C-, M-, or X-class flares from nonflaring instances.

3.4. Challenges and Limitations

Although our data products reveal some important correlations between PIL features and flare events, there are limitations in our approach. As mentioned in Section 2.3.1, we conduct a validation test on magnetograms to avoid severe projection effects when producing the PIL binary mask. This will directly cause partially or totally unexploited magnetograms by the detection algorithm, due to their proximity to the solar limbs. In addition, our data set is produced under the uniform parameter setting that we discussed in Section 2.3. Modification or optimization of certain thresholds might have tangible, significant effects both on the metadata and on the results of population separation or prediction experiments. Finally, using a uniform parameter setting, we treat PILs in the same magnetogram as a whole PIL set; thus, the PIL metadata reflect the overall performance of this PIL set. We did not attempt to further divide the PIL set into different subsets based on their attributes, such as size or enclosed flux. Such tasks, however, can be readily facilitated in future studies given this openly accessible PIL data set.

4. Conclusion

In considering the correlation between PIL and solar eruptive activity, using the primary features of PILs can be another important approach benefiting space weather forecasting. In this work, we present a novel and publicly available PIL data set, generated by line-of-sight and radial magnetograms, for heliophysicists and data scientists aiming to reveal and better understand the structure and temporal evolution of PILs. Our PIL data product not only includes time series features but also binary masks that are able to serve as inputs for PIL-related machine-learning models. Such binary masks are saved at the same spatial analysis as their corresponding magnetogram patches, so that users do not need to perform any extra image transformation in order to map them back to the original magnetogram. We also present three preliminary, exploratory analyses as examples of the very least that can be done with our data set—far more sophisticated learning or statistical tasks can be conceived and implemented readily. Even these initial analyses, however, either on the radial or the line-of-sight magnetic field component, demonstrate the portability and adaptability of our published, open-source data set. Hopefully the PIL data we make available will facilitate meaningful future studies of solar active regions and the eruptive activity thereof, exploiting both purely statistical and machine learning methods.

This project has been supported in part by funding from the Division of Advanced Cyber infrastructure within the Directorate for Computer and Information Science and Engineering, the Division of Astronomical Sciences within the Directorate for Mathematical and Physical Sciences, and the Division of Atmospheric and Geospace Sciences within the Directorate for Geosciences, under NSF award No. 1931555. It was also supported in part by funding from the Heliophysics Living with a Star Science Program, under NASA award No. NNX15AF39G.

Appendix A
Preliminary Exploratory Analyses on *B*_LoS Magnetograms

In this appendix, we present our preliminary exploratory analyses using features of our PIL data set derived from SDO/HMI line-of-sight (*B*_LoS) magnetograms.

A.1. Exploratory Analysis 1: Flaring versus Eruptive Active Regions

Similar to the analysis for the PIL data set derived from radial magnetic field (*B*_r) magnetograms, we label the flaring (associated with at least one C+ class solar flare) and eruptive active region (associated with at least one C+ class eruptive flare). In total, we have the same amount of active regions counted in our data set generated from *B*_LoS magnetograms (661 flaring active regions and 168 eruptive active regions). Again, we only take into consideration those active regions that last for more than 24 hr as well as those that passed the validation quality check.

As shown in Figure 6, we notice a similar distribution among all selected features. As the daily number of associated flares increases, the number of eruptive active regions also rises, with the majority of these active regions associated with two or more flares per day. We can also see that between the statistical analyses of features, there are certain thresholds that can distinguish between the flaring and eruptive active regions. For example, when the average number of PILs exceeds 10 for the entire lifespan of an active region, around 91.9% of the active regions are marked as eruptive. As the standard deviation of the number of PILs exceeds 4, around 86% of the marks are considered eruptive active regions.

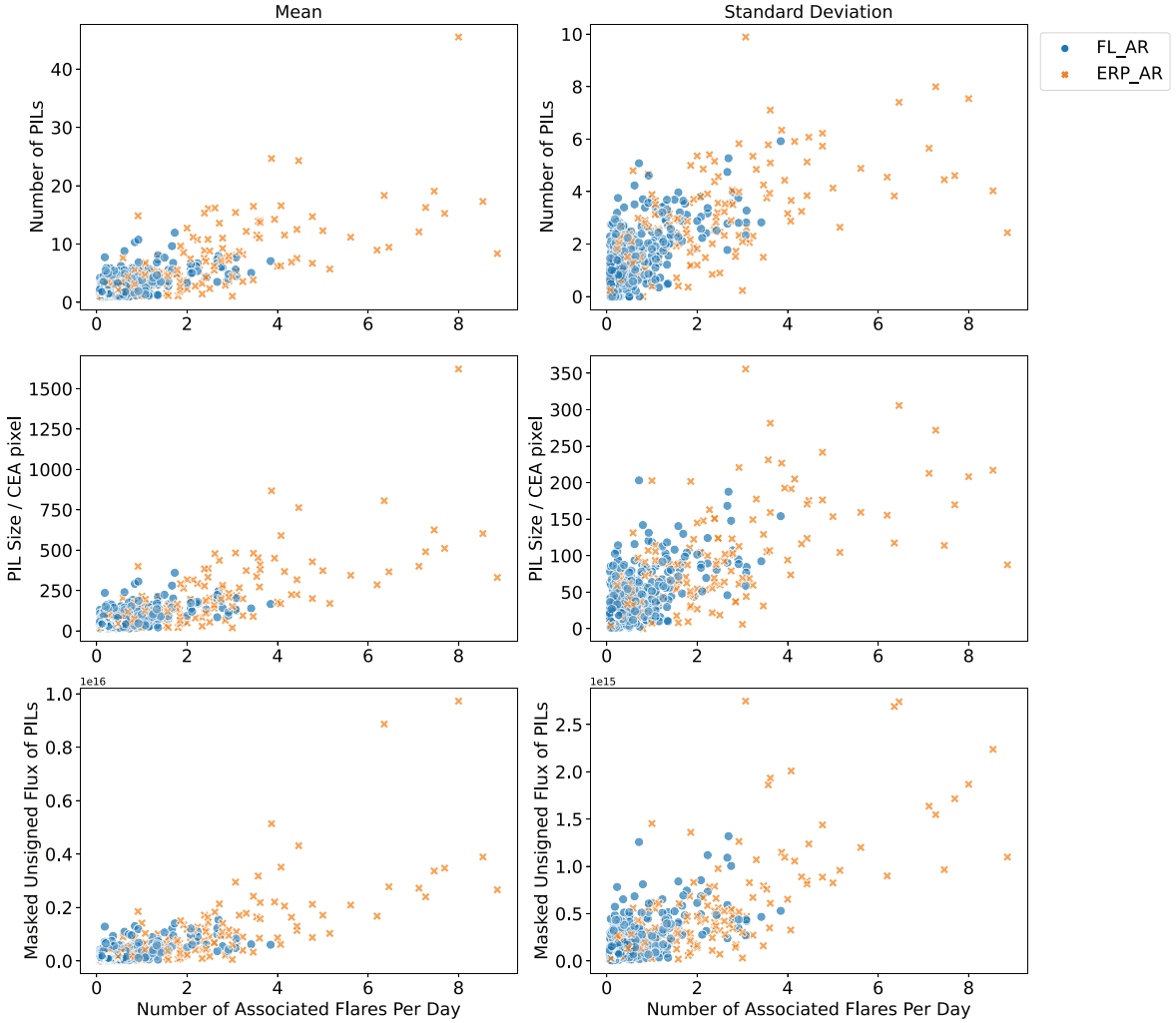

Figure 6. Mean and standard deviation scatter plots of three features (i.e., number of PILs, PIL size, and masked unsigned flux of PIL) derived from the entire flaring active region time series in regard to their relationships with number of associated flares per day.

A.2. Exploratory Analysis 2: Different Observational Time Period prior to Flare

In the second analysis, we calculate the statistical measurements and present the results using boxplots shown in Figure 7 for the means and Figure 8 for the standard deviations. Although there are major overlaps across the distributions of different classes of flares, we see a general decreasing trend across distribution characteristics that is more evident for the median values. This illustrates a consistency with the experimental results we showed previously. The higher values of our selected features (i.e., number of PILs, PIL size, and masked unsigned flux of PIL) are closely associated with relatively larger flares. Observing from Figure 7, we can see that the maximum and minimum of M-class distributions in the last feature (i.e., masked unsigned flux of PIL) exceed X-class distribution and the same for the first feature (i.e., number of PILs) in Figure 8.

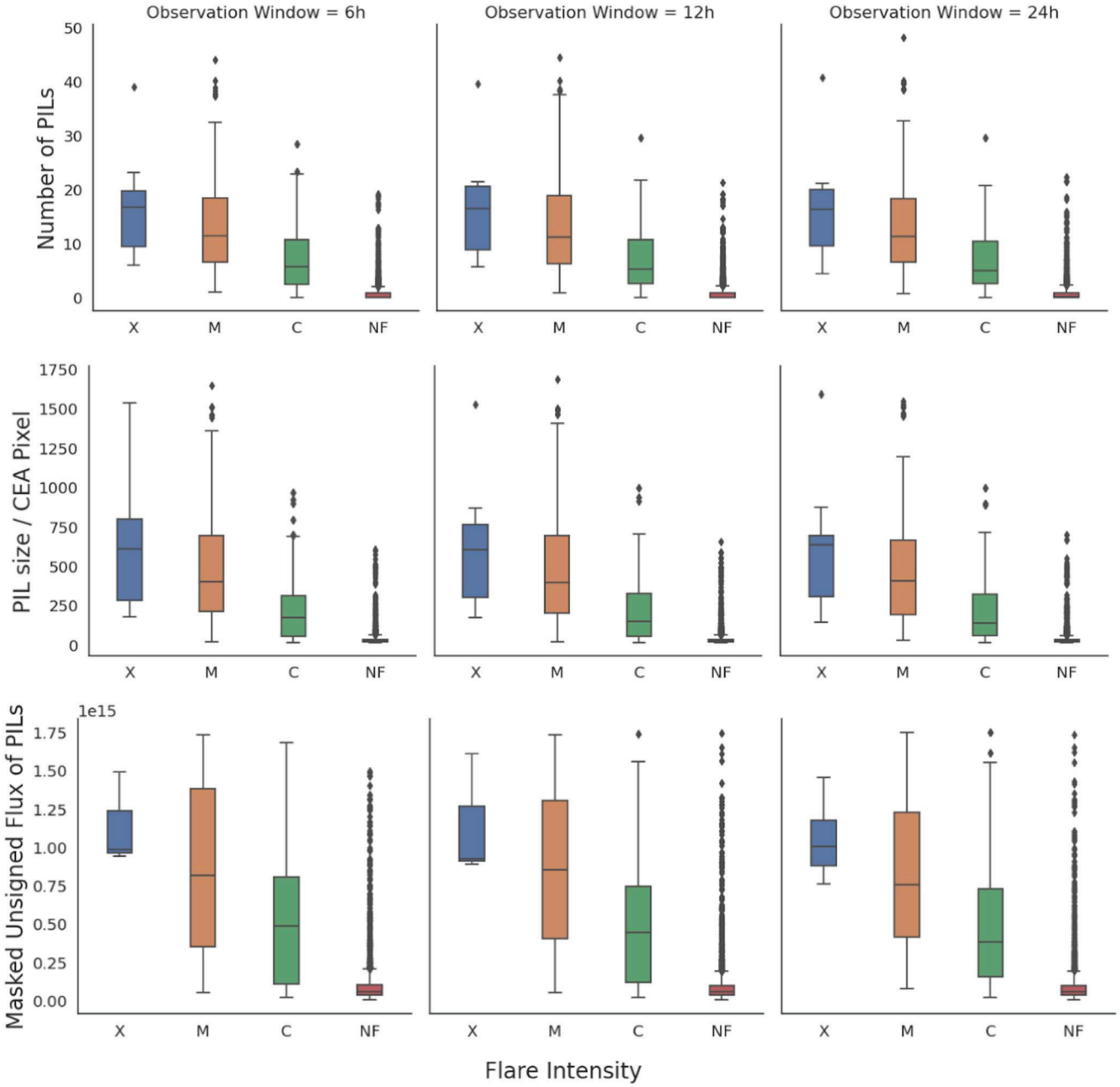

Figure 7. Box plots of means between selected PIL features prior to flaring and nonflaring instances based on GOES flare intensity level for 6, 12, and 24 hr time intervals.

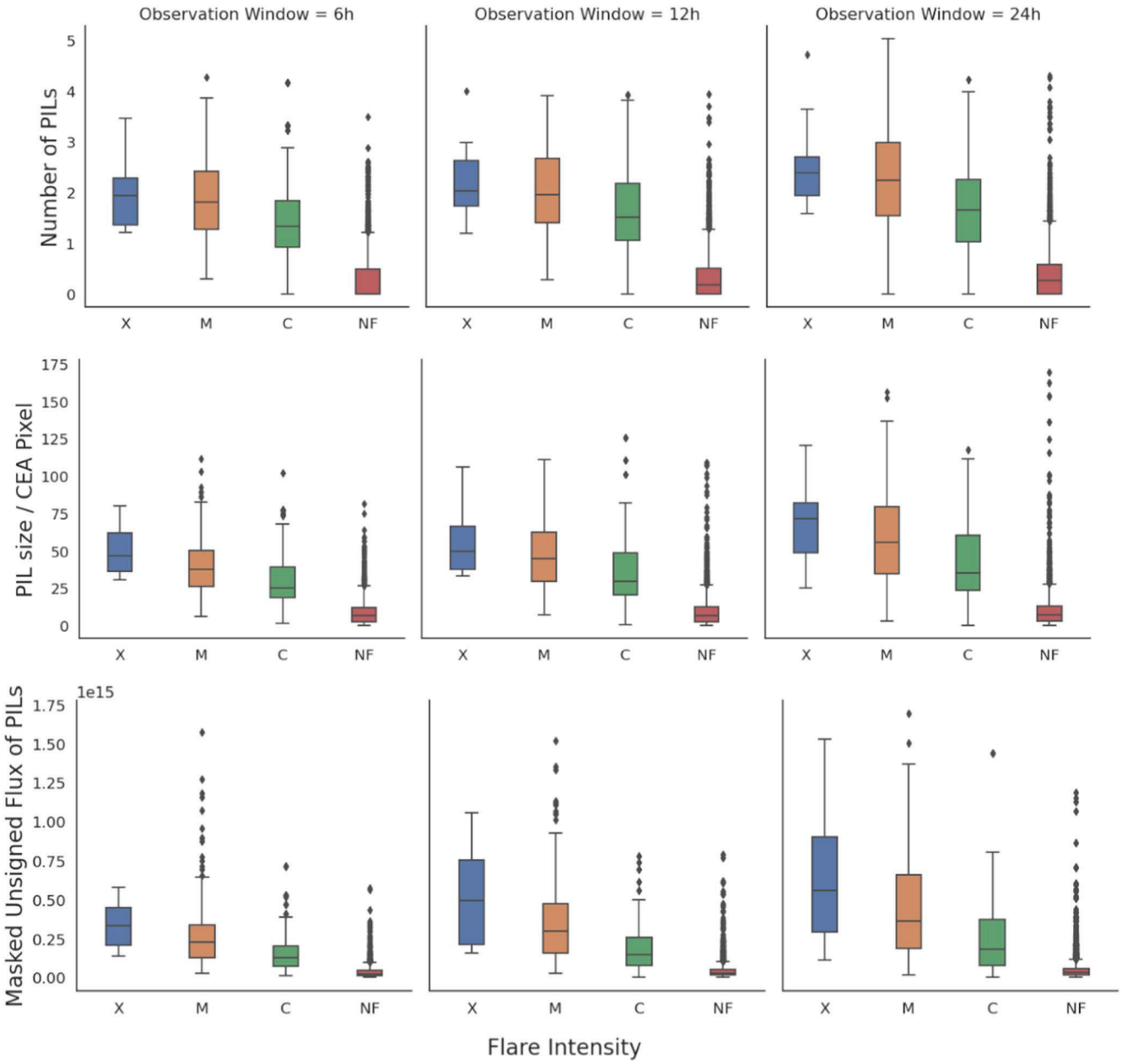

Figure 8. Box plots of standard deviations between selected PIL features prior to flaring and nonflaring instances based on GOES flare intensity level for 6, 12, and 24 hr time intervals.

A.3. Exploratory Analysis 3: Similarity and Discrimination between Flare Intensities

In our third analysis, we examine the characteristics of PILs and their discriminations across different flare intensity levels. Similar to the analysis for our PIL data set generated from Br magnetograms, we have separated flares of GOES classes into five subcategories: C1–C4 for C1.0- through C4.9-class flares, C5–C9 for C5.0- through C9.9-class flares, M1–M4 for M1.0- through M4.9-class flares, M5–M9 for M5.0- through M9.9-class flares, and X1+ for flares X1.0 and above. For each category, we examine the distributions of statistical time series features (i.e., mean and standard deviation derived with a 12 hr observation window) for the three previously selected PIL features: number of PILs, PIL size, and masked unsigned flux of PILs. We again create histograms of 20 equal bins for each statistical feature-parameter combination, where the same bin ranges were applied to all subcategories per combination. Then the dissimilarities are evaluated by using the discrete version of the Bhattacharyya distance. Higher dissimilarity represents more separability between the two density distributions.

Our results are summarized in Figure 9 as a set of heat maps. The diagonal plots provide the histogram distribution for the five subcategories. Our remarks are then summarized as follows:

1. Histograms in the diagonal often present a normal (i.e., right-skewed) distribution. However, in M1–M4 class histograms (mean series, in particular), we observe outliers toward higher values.
2. The maximum distance is observed between lowest (C1–C4) and highest (X1+) intensity classes consistently across feature-parameter combinations. The highest distance is achieved for the mean of PIL size and the

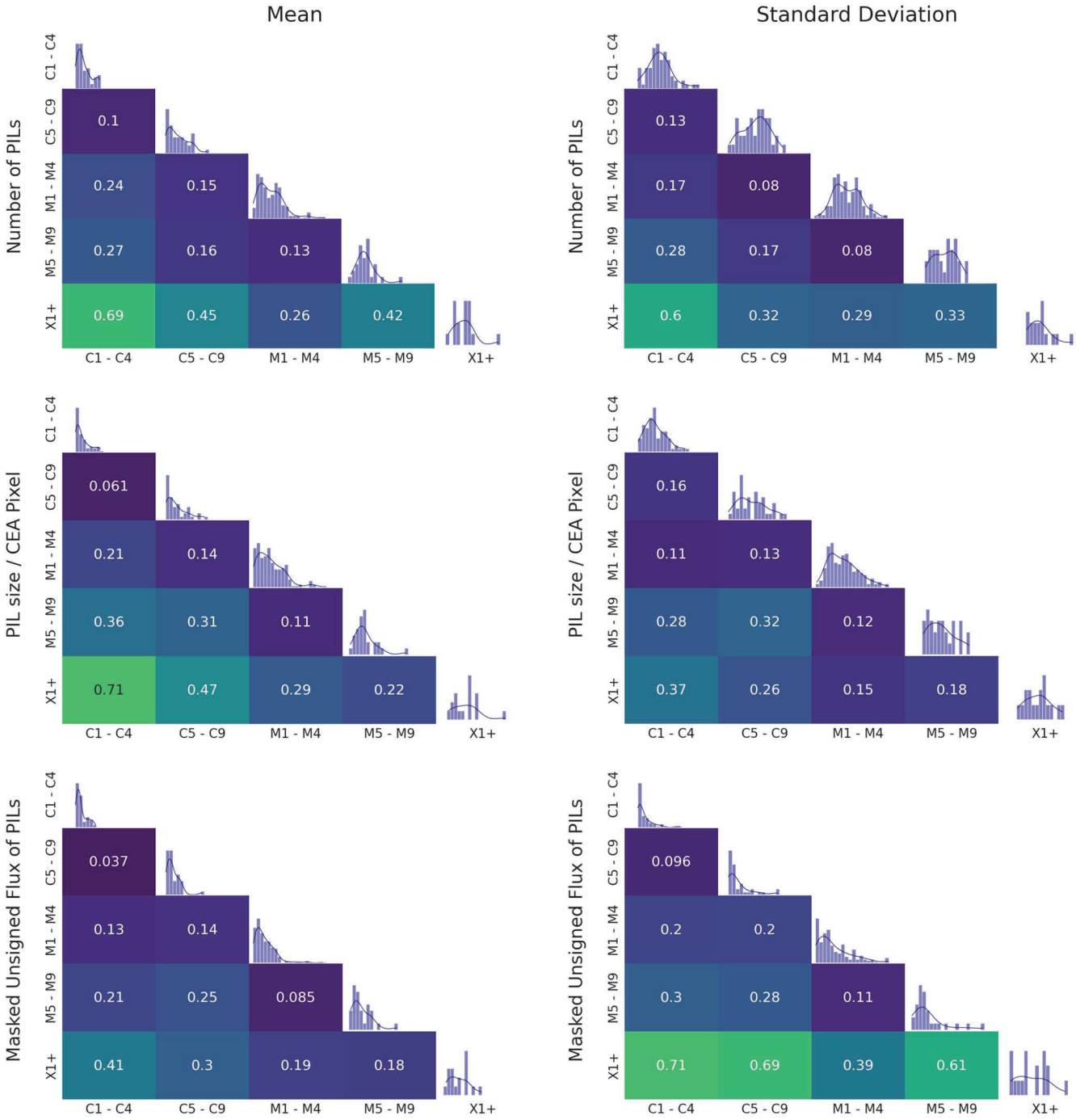

Figure 9. Heat maps for comparing similarity between different flare intensities by using the Bhattacharyya distance. The histogram on top of each column shows how flares are distributed within each category.

standard deviation of masked unsigned flux (for C1–C4 and X1+ subcategories).

3. In general, neighboring classes are least separable, as expected. Among these, C-class flares (i.e., C1–C4 and C5–C9 subclasses) and M-class flares (i.e., M1–M4 and M5–M9 subclasses) show the highest histogram similarity. Our results suggest that for separating X1+ flares, the mean of the masked unsigned flux of PILs parameter is the most viable univariate feature.
4. The mean of the time series appears generally more effective than the standard deviation in terms of histogram separability. There are a few exceptions to this: for example, the standard deviation of PIL size is slightly better than the mean (0.06 versus 0.16) when separating between C1–C4 and C5–C9.

Appendix B
Exploratory Analysis of Largest Flares on *B*_r Magnetograms

In this section, we conduct a further analysis of our *B*_r data set by taking only the largest flare from each active region, considering that the potential reason for some overlapping as seen in the previous study might be due to the massive amount of C-class flares. This means that the analysis can be also affected by the M- or C-class flares as a region may have already given, or is about to give, an X-class flare. Using only the largest flares will be one way to reduce the amount of noise. The results are presented in Figure 10 for the means and Figure 11 for the standard deviations. Compared to applying all flares from all active regions, as expected, we see less overlapping between different classes and the means of each

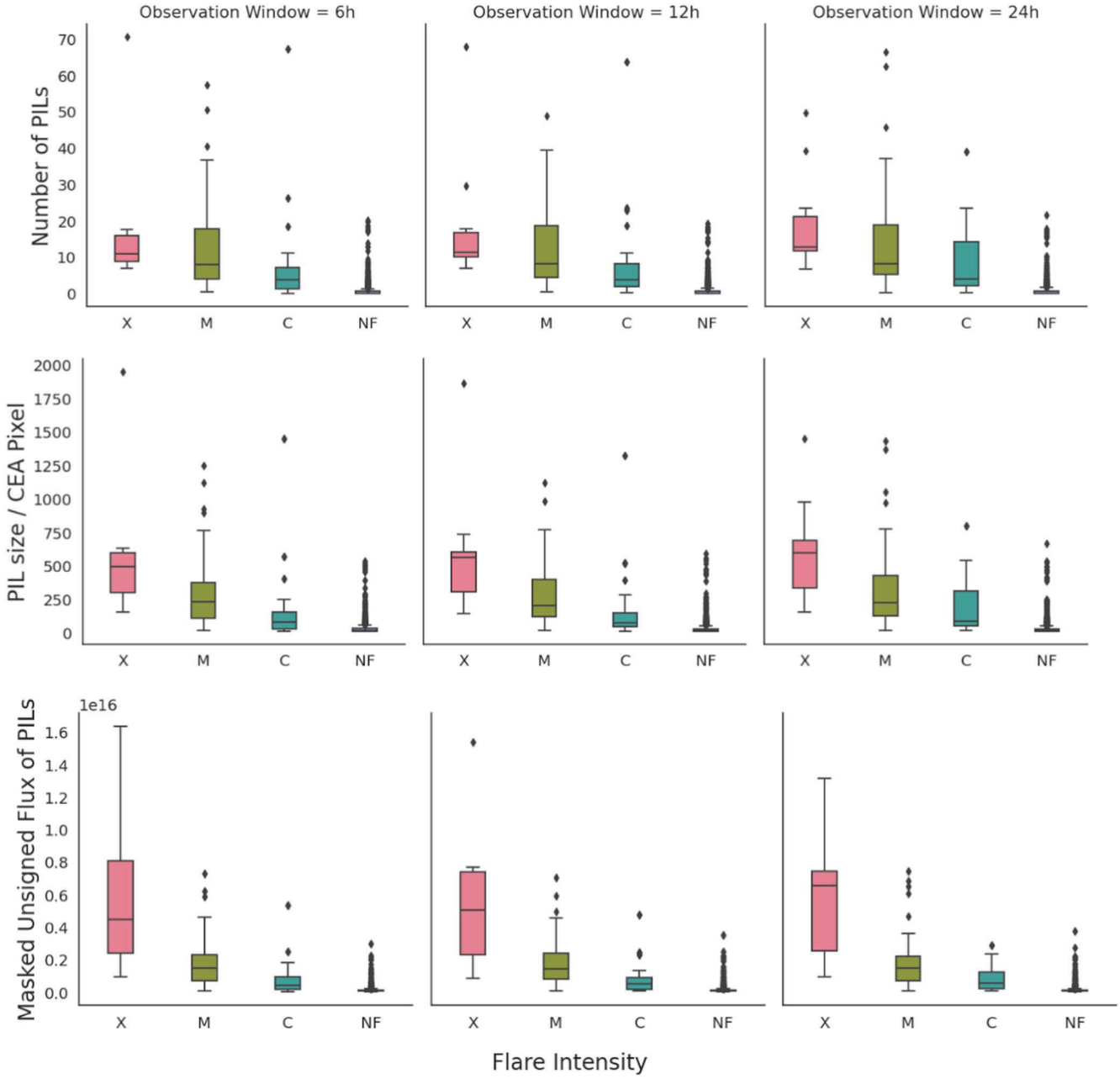

Figure 10. Box plots of means between selected PIL features prior to flaring and nonflaring instances based on GOES flare intensity level for 6, 12, and 24 hr time intervals with the largest flare from each active region in the radial magnetic field magnetograms.

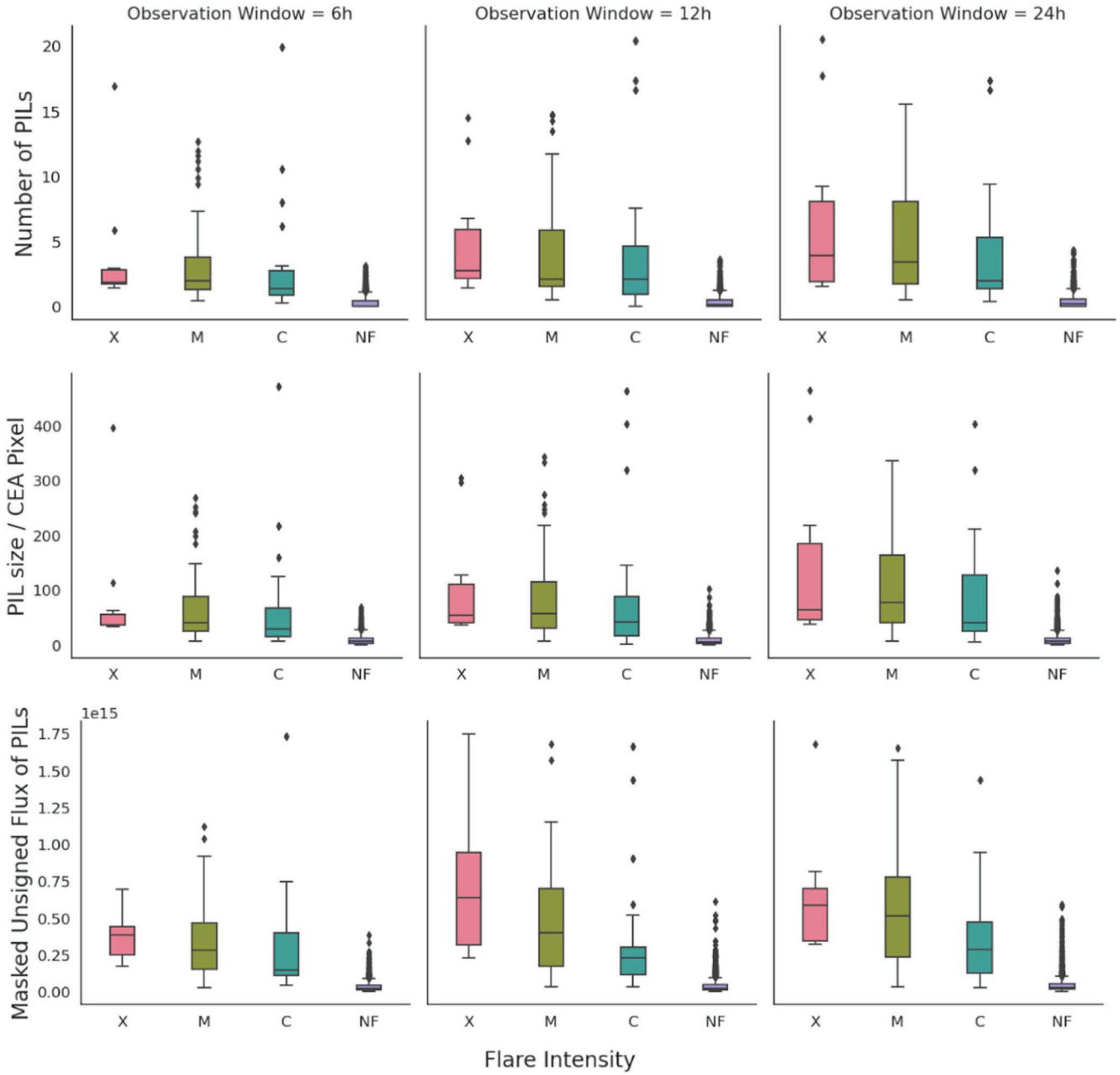

Figure 11. Box plots of standard deviations between selected PIL features prior to flaring and nonflaring instances based on GOES flare intensity level for 6, 12, and 24 hr time intervals with largest flare from each active region in the radial magnetic field magnetograms.

boxplot have a clearer trending from X to C classes. In such cases, we have a better understanding of PIL features and how they connect to flares of different classes.

ORCID iDs

Anli Ji https://orcid.org/0000-0002-1551-2370
Manolis K. Georgoulis https://orcid.org/0000-0001-6913-1330
Petrus C. Martens https://orcid.org/0000-0001-8078-6856
Rafal A. Angryk https://orcid.org/0000-0001-9598-8207
Berkay Aydin https://orcid.org/0000-0002-9799-9265

References

Antiochos, S. K., DeVore, C. R., & Klimchuk, J. A. 1999, ApJ, 510, 485
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Cai, X., Aydin, B., Ji, A., Georgoulis, M. K., & Angryk, R. 2020, in 2020 IEEE Int. Conf. on Big Data, ed. X. Wu et al. (Piscataway, NJ: IEEE), 4175
Campi, C., Benvenuto, F., Massone, A. M., et al. 2019, ApJ, 883, 150
Canny, J. 1986, ITPAM, PAMI-8, 679
Cicogna, D., Berrilli, F., Calchetti, D., et al. 2021, ApJ, 915, 38
Falconer, D. A., Moore, R. L., Barghouty, A. F., & Khazanov, I. 2014, SpWea, 12, 306
Georgoulis, M. K., Nindos, A., & Zhang, H. 2019, RSPTA, 377, 20180094
Gopalswamy, N., Yashiro, S., Michalek, G., et al. 2009, EM&P, 104, 295
Hagyard, M. J., Moore, R. L., & Emslie, A. G. 1984, AdSpR, 4, 71
Heyvaerts, J., Priest, E. R., & Rust, D. M. 1977, ApJ, 216, 123
Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483
Ji, A., & Aydin, B. 2022, Flare to CME Association Integration, V1, Harvard Dataverse doi:10.7910/DVN/WSEY4T
Jones, H. P. 2004, in Knowledge-based Intelligent Information and Engineering Systems, ed. M. G. Negoita (Berlin: Springer), 433
Kailath, T. 1967, ITCom, 15, 52

Korsós, M. B., Yang, S., & Erdélyi, R. 2019, JSWSC, 9, A6
Mandelbrot, B. 1985, PhyS, 32, 257
Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634
Ming-Kuei, H. 1962, ITIT, 8, 179
Mora, C., & Kwan, A. 2000, Cement Concrete Res., 30, 351
Murray, S. A., Guerra, J. A., Zucca, P., et al. 2018, SoPh, 293, 60
Padinhatteeri, S., Higgins, P. A., Bloomfield, D. S., & Gallagher, P. T. 2015, SoPh, 291, 41
Park, S.-H., Guerra, J. A., Gallagher, P. T., Georgoulis, M. K., & Bloomfield, D. S. 2018, SoPh, 293, 114
Sadykov, V. M., & Kosovichev, A. G. 2017, ApJ, 849, 148
Serra, J. 1983, Image Analysis and Mathematical Morphology (New York: Academic)
Severny, A. B. 1964, ARA&A, 2, 363
Snyder, J. P. 1987, Map Projections: A Working Manual, Professional Paper 1395 US Geological Survey
Sun, H., Manchester, W., & Chen, Y. 2021, SpWea, 19, e2021SW002837
Toriumi, S., & Wang, H. 2019, LRSP, 16, 3
van Ballegooijen, A. A., & Martens, P. C. H. 1989, ApJ, 343 971
van Driel-Gesztelyi, L., & Green, L. M. 2015, LRSP, 12, 1
Vlahos, L., & Georgoulis, M. K. 2004, ApJL, 603, L61
Volobuev, D. M. 2011, Strong Polarity-Inversion Line Preceding an X-Class Flare, http://publications-dv.narod.ru/SPIL-DV3.pdf
Wang, J., Liu, S., Ao, X., et al. 2019, ApJ, 884, 175
Wang, J., Zhang, Y., Webber, S. A. H., et al. 2020, ApJ, 892, 140
Woodcock, N. H. 1977, GSAB, 88, 1231
Zirin, H., & Liggett, M. A. 1987, SoPh, 113, 267