Draft version October 16, 2022
Typeset using LaTeX preprint style in AASTeX63

# Properties of Flare-Imminent versus Flare-Quiet Active Regions from the Chromosphere through the Corona II: NonParametric Discriminant Analysis Results from the NWRA Classification Infrastructure (NCI)

K.D. Leka,[1,2] Karin Dissauer,[1] Graham Barnes,[1] and Eric L. Wagner[1]

[1]NorthWest Research Associates, 3380 Mitchell Lane, Boulder, CO 80301 USA
[2]Institute for Space-Earth Environmental Research, Nagoya University,
Furo-cho Chikusa-ku, Nagoya, Aichi 464-8601 JAPAN

(Received; Revised; Accepted)

Submitted to ApJ

## ABSTRACT

A large sample of active-region-targeted time-series images from the Solar Dynamics Observatory / Atmospheric Imaging Assembly, the AIA Active Region Patch database ("AARPs", Paper I: Dissauer et al. (2022c)) is used to investigate whether parameters describing the coronal, transition region, and chromospheric emission can differentiate a region that will imminently produce a solar flare from one that will not. Parametrizations based on moment analysis of direct and running-difference images provide for physically-interpretable results from nonparametric discriminant analysis. Across four event definitions including both 24 hr and 6 hr validity periods, 160 image-based parameters capture the general state of the atmosphere, rapid brightness changes, and longer-term intensity evolution. We find top Brier Skill Scores in the 0.07 – 0.33 range, True Skill Statistics in the 0.68 – 0.82 range (both depending on event definition), and Receiver Operating Characteristic Skill Scores above 0.8. Total emission can perform notably as can steeply increasing or decreasing brightness, although mean brightness measures do not, demonstrating the well-known active-region-size/flare-productivity relation. Once a region is flare productive, the active-region coronal plasma appears to stay hot. The 94 Å filter data provides the most parameters with discriminating power, with indications that it benefits from sampling multiple physical regimes. In particular, classification success using higher-order moments of running difference images indicate a propensity for flare-imminent regions to display short-lived small-scale brightening events. Parameters describing the evolution of the corona can provide flare-imminent indicators, but at no preference over "static" parameters. Finally, all parameters and NPDA-derived probabilities are available to the community for additional research.

Corresponding author: K. D. Leka
Corresponding author: Karin Dissauer
leka@nwra.com

*Keywords:* methods: statistical – Sun: flares – Sun: corona – Sun: chromosphere

## 1. INTRODUCTION

In Dissauer et al. (2022c, hereafter Paper I) we briefly introduce the goal for this study: to quantitatively characterize the brightness distributions, their temporal variations and implied kinematics, and eventually a more complete physical state of the chromosphere and corona, for two populations of solar active regions: those that are flare-productive on specified time-scales *vs.* those that are not. We are addressing this goal with a large sample of data from the Atmospheric Imaging Assembly (AIA; Lemen et al. 2012) onboard the Solar Dynamics Observatory (SDO; Pesnell et al. 2012, see Section 2). There has not yet been such a characterization in the context of flare productivity. The approach we invoke explicitly avoids focusing on "pre-flare"-specific phenomena, and instead examines more general behaviors.

Recently, the dominant use of large-sample coronal image data in the context of solar energetic phenomena has been for machine learning tools to try and predict solar flares (Nishizuka et al. 2017; Jonas et al. 2018; Alipour et al. 2019, although see (Krista & Chih 2021)). Generally, these statistical tools have not yet provided "interpretable" results in terms of a physics-based outcome, although they have demonstrated some added classification success when combining coronal data with, *e.g.*, photospheric magnetic field data from the Helioseismic and Magnetic Imager (HMI; Scherrer et al. 2012; Hoeksema et al. 2014; Bobra et al. 2014).

Case-study analyses of the pre-event solar corona have found evidence of loop formation, energization and increased dynamic behavior ("crinkles"; Sterling & Moore 2001a; Joshi et al. 2011; Sterling et al. 2011; Imada et al. 2014, and references therein), an increase in chromospheric non-thermal velocities and high blueshifts (Cho et al. 2016; Harra et al. 2013; Woods et al. 2017; Seki et al. 2017), very localized chromospheric heating (Li et al. 2005; Bamba et al. 2014), and coronal dimming (Imada et al. 2014; Zhang et al. 2017; Qiu & Cheng 2017) in the hours prior to energetic events.

The present study attempts to do for the solar corona and chromosphere what was done for the photosphere in a previous series of papers (Leka & Barnes 2003a,b; Barnes & Leka 2006; Leka & Barnes 2007; Leka et al. 2018): test the ability to statistically differentiate between flare-quiet and flare-imminent active regions through analysis of photospheric magnetic field data. Here we begin to test the same question but with a focus on the chromosphere, transition region, and corona. Guided by the previous series of papers, we use here active regions as defined by the HMI Active Region Patches (HARPs: Hoeksema et al. 2014) but now use time-series images of the upper solar atmosphere in the UV and EUV (Section 2.1; see also Paper I). We introduce human-constructed parametrizations (Section 3.1) designed to provide insights into the physical state of the upper atmosphere in a manner parallel to what the "SHARP parameters" (Bobra et al. 2014) and especially the extended parameter list examined in Leka & Barnes (2007); Leka et al. (2018) provide for the photosphere (see also Georgoulis et al. 2021, and references therein). Without focusing on forecasting *per se*, here we extend insights gained by prior case studies to a large sample, to statistically test (Section 3.2) whether we can differentiate the state of active region atmospheres that are flare-imminent from those that are not.

Employing a large sample size provides a broad picture not only of the standard workings of the corona over all sizes and activity levels of active regions, but to what extent there is such a thing as standard workings. In other words, what is important for our understanding of the Sun is not only the mean of some characteristics, but the more nuanced nature of the distributions of those characteristics, their degree of overlap, *etc*. Here we quantify some characteristic behaviors between defined groups, setting empirically-derived standards to which models may then need to speak.

## 2. THE DATA

The observational data used in this study are described in this section, both the AIA timeseries data (Section 2.1) and the data used (Section 2.2) to construct the solar flare event lists for analysis (Section 3.2.1).

### 2.1. *The AIA Active Region Patches (AARPs)*

The AIA Active Region Patches (AARPs) database is described in full in Dissauer et al. (Paper I; 2022c). Broadly speaking, they consists of curated UV- and EUV-image timeseries counterparts to the photospheric magnetic field time-series data deployed in Leka et al. (2018).

The primary data source used in constructing the AARPs is *SDO*/AIA, supplemented with meta-data from the Helioseismic and Magnetic Imager (HMI; Scherrer et al. 2012; Hoeksema et al. 2014) `hmi.Mharp_720s` series. The latter provides the coordinates and bounding-box of the HMI Active Region Patches (HARPs; Hoeksema et al. 2014), which are the basis for defining the areas extracted from the AIA full-disk images. Of note, however, the AARP boxes are larger by 20% than the HARP definitions in order to accommodate the larger projected extent of the 3-D coronal structures, especially when a region is located near a limb, and the bounding-box is extended further in the limb-direction to include the AR loops (see Paper I for details). There is no spatial binning applied to the images.

For each numbered HARP on any particular day, there is one corresponding AARP consisting of seven hourly samples each containing 13 min of data sampled at 72 s (11 images), across each of eight AIA bands. To match the database of HMI vector magnetic field extractions already in place at NWRA, the seven hourly samples span 15:48 TAI – 21:48 TAI. FITS files are produced for each of seven EUV filters (94, 131, 171, 193, 211, 304, and 335 Å), and the UV 1600 Å. This approach provides information on both short-term and longer-term evolution of all magnetic patches at chromospheric, transition region, and coronal heights and temperatures. The NWRA AARP database, which is available at the Solar Data Analysis Center (Dissauer et al. 2022b), is summarized in Table 1; here the number of samples is the total number of AARP datasets available over the full date range. The AARPs provide the data for parametrization (Section 3.1), so the number of samples in Table 1 is the total sample size available for statistical analysis for the present study. There is no further down-selecting for AR size, complexity, location, or activity level.

Table 1. Summary of AARP Data Set

| Date Range | AARP Range | NOAA AR Range | Number of "AARP-Day" Samples | Archive Size |
|---|---|---|---|---|
| 06/2010 – 12/2018 | 36 – 7331 | 11073 – 12731 | 32,067 | ≈ 9.5 TB |

### 2.2. *GOES Data and Source for Event Lists*

The event lists are constructed following Leka et al. (2018), using events as recorded by NOAA using the Geostationary Operational Environmental Satellite X-Ray Sensor ("GOES"/XRS Garcia 1994). The dataset used is consistent with regards to flux calibration (Viereck & Machol 2017; Machol 2022). Only those events associated with NOAA-assigned Active Regions are included. Flare lists based on GOES 1–8A peak emission from the GOES/XRS sensors are available through either the National Center for Environmental Information (NCEI) or by way of the "edited event lists" from NOAA/Space Weather Prediction Center. In the present study we used the latter by which to construct the event lists used (see Section 3.2.1).

## 3. ANALYSIS

The question posed here is, "for solar active regions, are flare-imminent epochs distinguishable from flare-quiet epochs on the basis of chromospheric and coronal emission and kinematics?" Specifically we ask this using UV and EUV intensity images and HMI-defined active regions, without the added benefit of spectroscopy (Panos & Kleint 2020), but with the explicit use of time-series analysis (Cinto et al. 2020) in order to enhance physical interpretation of the results. We answer the question through statistical classification, multiple event definitions, and quantitative metrics to evaluate how well the samples can differentiate the two populations.

### 3.1. *Parametrization*

Parametrization allows both spatial and temporal information to be summarized succinctly and in a manner conducive to physical interpretation upon statistical analysis. Moment analysis through the fourth moment is used on the spatially-sampled target x: mean $\mu(x)$[1], standard deviation $\sigma(x)$, skew $\varsigma(x)$, and kurtosis $\kappa(x)$. The lower-order moments capture bulk differences whereas the higher-order moments are much more sensitive to subtle differences in distribution wings, but are also more susceptible to errors when image sizes are small. The odd moments detect offsets or asymmetries as related to a normal distribution, whereas the even moments are sensitive to deviations in width or peakedness. Previous research of magnetic field distributions (Leka & Barnes 2003b; Barnes & Leka 2006; Leka & Barnes 2007; Barnes et al. 2007; Leka et al. 2018) shows the power of 3rd and 4th-order moments to capture subtle differences in distribution *tails* that can signal significant, but very localized, changes – such as from a small emerging flux region.

The moment-analysis parametrizations produce a selection of *intensive* variables that do not scale directly with active region size (Welsch et al. 2009); these are complemented by *extensive* parameters (such as totals over the field of view), which *do* scale with region size. It is important to note that the moment analysis is not intended to provide a basis for image decomposition (Raboonik et al. 2017) and as such, while the resulting parametrizations may not be unique, they readily allow interpretation of the image intensity behavior. The parametrization is applied to the images by themselves, what we call the "direct" images ("I"), as well as the running-difference images ("ΔI"). For this analysis, the parameters target the following (for each wavelength separately, indicated by "_*"):

- The total brightness of an image, $\Sigma(I_{\square})$, and of the running-difference image $\Sigma(\Delta I_{\square})$.

[1] To avoid confusion, we use here $\mu(x)$ for mean(x) which breaks with our previous use of $\bar{x}$; we also refer explicitly to the cosine of the observing angle $\cos(\theta)$ without invoking $\mu$ in that context.

- The moments of the brightness distribution $M(I_{\square})$ which summarizes the mean $\mu(I_{\square})$, standard deviation $\sigma(I_{\square})$, skew $\varsigma(I_{\square})$, and kurtosis $\kappa(I_{\square})$.
- The moments of the running-difference image distributions $M(\Delta I_{\square})$, which summarizes individually the mean $\mu(\Delta I_{\square})$, standard deviation $\sigma(\Delta I_{\square})$, skew $\varsigma(\Delta I_{\square})$, and kurtosis $\kappa(\Delta I_{\square})$.
- The cosine of the central observing angle cos(θ); this is essentially used as a control since flare activity should not have preferred locations.

In all, 80 base parameters are defined plus the observing angle (the same for all wavelengths). A parameter X is computed for each of 11 images (or each of 10 running-difference image) within the 13-minute sample (see Figure 1). The average and standard deviation of these 11 (10) is assigned to the mid-time (the ":48") that matches the hmi.Mharp_720s data (see Figure 2, top panels), the standard deviation being used as an estimate of the uncertainty of that parameter over the 13 min. This procedure is performed for each of the 7 hourly samples (see Figure 2, bottom panels).

A parameter X's "static" state and its temporal behavior dX/dt are finally described using the slope and intercept (at the last data sample's central time (using T_REC), 21:48 TAI) of a linear fit over the 7 hourly samples (Figure 2, bottom panels), following the magnetic field analysis in Leka et al. (2018). Of note, parameters that are by definition positive- or negative- definite are limited in the "static" parameter to the appropriate sign; if the inferred value by the intercept of the fit does imply a crossing in sign, the returned parameter is set to 0.0. Data outages exist; at minimum, 2 data points are required, for which only the mean is returned as the static parameter, and the dX/dt is returned as a NaN. To fit the slope, we require a minimum of 3 data points. We have found that a linear fit is sufficient to describe the general behavior without over-fitting for short-timescale fluctuations. We (de-)weight the fits by the uncertainties at each time, and one or a few outlier data points rarely corrupt the linear fits, especially if they include large uncertainties. Flares occur during the data acquisition (Figure 3) but rarely do their influence persist more than 2–3 hr, and they are usually extremely variable on short timescales (resulting in large uncertainties in the hourly means of the parameters). As such, the linear fits generally all but ignore them. That being said, there exist "perfect storm" situations that will introduce outlier points. One example is 2016.01.20, AARP#6281 where two B-class flares occurred between 15:48–17:48 TAI, after which there was a data outage, so that only three points were available. The parameters for this AARP on this day were severely influenced (*e.g.* $d(\kappa(I_{131}))/dt$). This situation can influence both the static and dX/dt parameters, but the latter may be more susceptible. That being said, we have examined the frequency of such outliers and have found that they typically occur no more than 0.1% of the time, which should not influence the final metrics beyond that level.

Thus, the final number is 160 image-based parameters plus the cos(θ) variable, for 161 independent parameters to be analyzed. These parametrizations are chosen to be physically interpretable. For example, one can expect that the appearance of new bright loops will enhance overall brightness levels of, for example, 171A images ($\Sigma(I_{171})$) and the mean brightness levels ($\mu(I_{171})$), but also possibly produce a distinct positive skew in the associated running-difference images ($\varsigma(\Delta I_{171})$) as the new loops appear. The brightness of coronal structures can also change due to heating or cooling (Viall & Klimchuk 2012) especially for 171A. On the other hand, we could expect that increased kinematic activity such as enhanced loop motion *without* significant brightness enhancements or new structures appearing will be signaled by broader distributions in running-difference images *without* an accompanying increase in the total, mean, or skew.

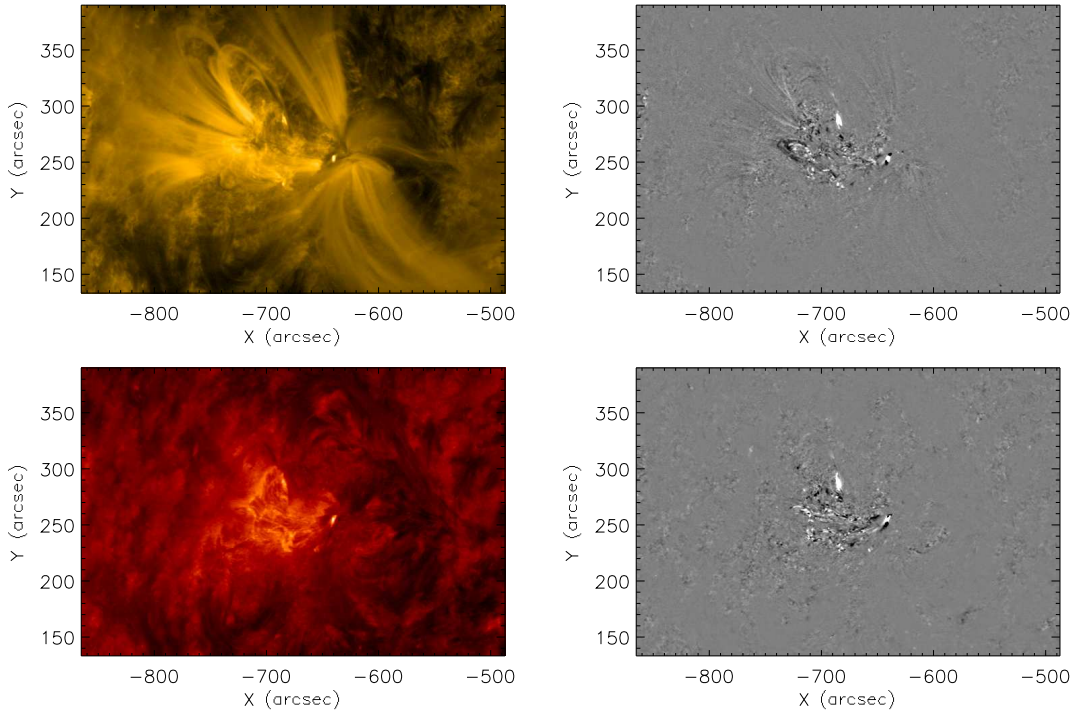



Figure 1. Demonstration of parametrizing AIA 171Å (top row) and AIA 304Å (second row) intensity (left) and running-difference (right) timeseries, for AARP 746, NOAA AR 11260; direct images: 2011-07-26T17:45:38Z, running-difference images are 2011-07-26T17:45:38Z - 2011-07-26T17:44:26Z. The running-difference variations are similar between the two but there is more structure in the AIA 171Å data that could provide additional information, or could be construed as noise by NCI. The procedure demonstration continues in Figure 2.

We do not, here, consider parameters that use base-difference or base-ratio analysis. The event definitions employed (Section 3.2.1) mean that the data sampling is agnostic as to the time of any event. Base-difference and similar approaches are most relevant when the base image refers to a known or specified state against which changes are measured (Plowman 2016). The running-difference images used here focus instead on evaluating the degree of variability of the atmosphere, by way of the intensity images, at the sampled times only.

### 3.2. *The NWRA Classification Infrastructure*

The NWRA Classification Infrastructure (NCI; Leka et al. 2018) is a well-established statistical classifier system based on Nonparametric Discriminant Analysis (NPDA). There are four components at work in this facility: the input parameters, the event definitions and event lists, the statistical package, and the evaluation metrics. We described the input parameters that will be used here, in Section 3.1, above. A general description of NCI is given in the referenced work, and below we describe the particulars as employed here.

#### 3.2.1. Event Definitions and Event Lists

The "event definition" includes all relevant characteristics to what defines "an event", such as details on timing, event size, event characteristics, *etc.* In this context, an event is when at least one flare above a specified threshold occurs during a specified validity period. A data point (*e.g.* a parameter

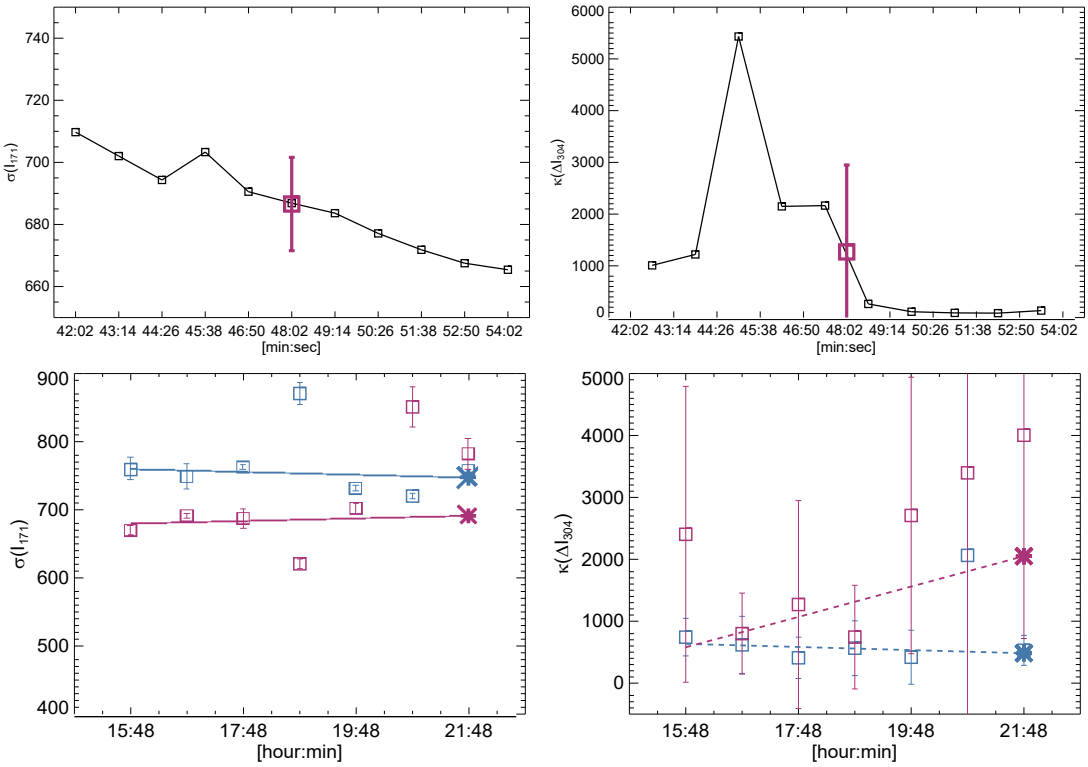



Figure 2. Following from Figure 1, the standard deviation of the brightness images of 171 Å is calculated for the 13 min (11 images) ("$\sigma(I_{171})$", left, top) centered at 2011.07.26 17:48 TAI, from which the mean and standard deviation are shown (thick point with error bar); these become the data points for each of the 7 samples covering 6 hours inclusive (left, bottom), from which the linear slope and last-data (21:48 TAI) intercept (thick asterisk) provide the final variables that are analyzed in NCI. Shown are the results for an M1.0+/24 hr "yes-event" sequence sample on 2011.07.26 (red) and a "no event" sample time period on 2011.07.25 (blue). The same sequence is shown for the kurtosis of the running-difference images of 304 Å("$\kappa(\Delta I_{304})$", right plots).

for one AARP) will be assigned to the flaring population in this case (Figure 3), and assigned to the flare-quiet population if no such events occurred. The assignments of AARPs to populations change according to the event definitions. We invoke NCI in its standard "prediction" mode which describes the timing definitions (see Figure 3). Specifically, there is no explicit coordination between the time of the events and the data acquisition time (as is the case for super-posed epoch analysis, *e.g.* Mason & Hoeksema 2010; Bobra & Couvidat 2015; Jonas et al. 2018).

The solar flare specific event definitions used here are described by (1) lower- and upper- peak intensity thresholds of peak GOES 1–8 Å flux (here upper-thresholds are set to infinity), (2) the validity period during which an event is predicted to occur, (3) the latency period that defines the interval between the end of the data and the beginning of the validity period. The event definitions considered here are summarized in Table 2. Some reflect standard definitions used for flare-prediction research, but some are more focused on shorter-term chromospheric and coronal behavior in the present context.

Of note, for M1.0+/24 hr and M1.0+/6 hr definitions, C-class and smaller flares are considered "non-events". Additionally, for all definitions, multiple qualifying flares within the validity window

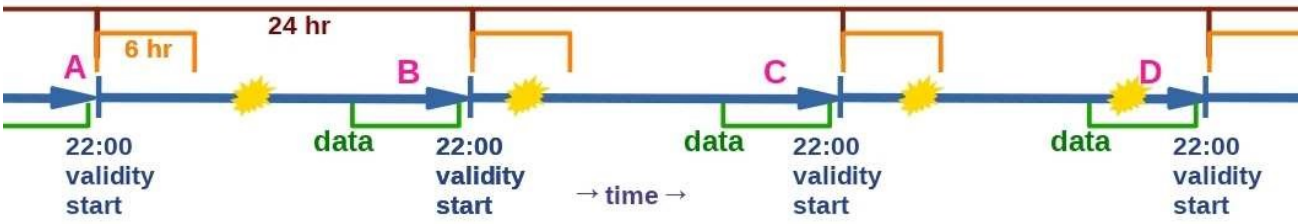



Figure 3. Schematic illustrating the relationship between the AARP data acquisition periods, the two validity periods invoked, and a few flares (events). Time proceeds to the right. 24 hr days are marked out by the blue arrows, with an implied start/stop time of 22:00 TAI. The data are acquired at the same time each day (green). The validity periods, both 6 hr (orange) and 24 hr (red) are indicated, all starting at 22:00 TAI. The first event (yellow graphic) would be a "yes-event" for the 24 hr validity period based on the data acquired by "A" but a "non-event" for the 6 hr period, whereas the data collected during "B" leads to a classification of the second flare as an "event" entry for both validity periods. The third and fourth events are classified according to the data collected in "C" even though it occurs during the "D" data collection, and would be designated an "event" for both the 6 hr and 24 hr definitions, even though there are two qualifying events for the latter within its validity period.

are considered together as a single positive event, so that the number of events may be smaller than the total number of flares during the period. Finally, a data point assigned to the "non-event" population may have previously or may subsequently flare – a "flare-quiet region" in the context of this analysis is a "flare-quiet epoch", or a time of no events, regardless of past or future activity.

One difference from earlier work on magnetic field-based analysis (Leka et al. 2018) is the start time for the validity periods. We matched the AARPs to the HMI-based database already in place (see Paper I). That database was constructed with anticipation to the delay in acquiring the near-real-time vector data for a true forecasting system that would produce forecasts starting at 00:00 UT (Leka et al. 2018). We have no such constraints here except the desire to match the HMI dataset.

Hence, the start time of the validity periods moved to 22:00 TAI for all event definitions. For the "24 hr" definitions, the validity time then runs from 22:00 TAI the day of the data acquisition, to 21:59:59 TAI the next day; in the case of the "6 hr" definitions, it runs from 22:00 TAI the day of the data acquisition to 03:59:59 TAI the next day. The "6 hr" definitions thus have significantly smaller event sample sizes, but the analysis becomes closer to "precursor" parameter evaluation.

### 3.2.2. NonParametric Discriminant Analysis

Table 2. Event Definition Summary

| Label | GOES lower limit $10^{-6}$ W m$^{-2}$ | Validity Period hr | Latency Period hr | # Events, (Event Rate R) |
|---|---|---|---|---|
| C1.0+/24 hr | 1.0 | 24 | 0.2 | 2752 (0.086) |
| M1.0+/24 hr | 10.0 | 24 | 0.2 | 450 (0.014) |
| C1.0+/6 hr | 1.0 | 6 | 0.2 | 1262 (0.039) |
| M1.0+/6 hr | 10.0 | 6 | 0.2 | 155 (0.005) |

Discriminant Analysis (DA) in general classifies input as belonging to one of two (or more) populations by dividing parameter space into two regions based on where the probability density of one population (*e.g.* flare-imminent regions) exceeds the other (*e.g.* not flare-imminent regions) so as to best separate the two samples. Discriminant Analysis does not simply look for correlations; a statistical classifier such as DA or Random Forest (Breiman 2001) divides parameter-space from samples of known populations, in the same mathematical "spirit" as machine-learning algorithms.

In NonParametric Discriminant Analysis (NPDA), no assumptions are made about the functional form of the distributions; instead, the probability density function is estimated directly from the data. Since it was described in Leka et al. (2018), we have added the capability of using adaptive kernel density estimation to NCI. This technique, used here, starts with a pilot density estimate from the Epanechnikov kernel and a fixed smoothing parameter determined by reference to a standard distribution (normal in this case; Silverman 1986; Leka & Barnes 2007), which works well for sufficiently large sample sizes, but tends to under-smooth the tails of a distribution and over-smooth the peak. This pilot density estimate is then used to estimate local bandwidth factors which determine the local width of the Epanechnikov kernel in combination with an overall sensitivity parameter, taken here to be $\alpha = 0.5$.

Although NCI with NPDA can be used for multi-variable analysis (multiple parameters simultaneously creating a higher-dimension parameter-space), we focus here on single-variable NPDA and strive for statistically-significant sample sizes for each event definition (Section 3.2.1) and a first-look set of results that can be physically interpretable. Example density functions and NPDA boundaries are given for select parameters in Figure 4, and discussed in Section 4, below.

NCI generates probabilities that a datapoint will belong to one or the other population based on the ratio of probability density function estimates from the samples plus the populations' prior probabilities. Note that as described in Leka et al. (2018), NCI treats "null" data and "bad" data differently. Additionally, in cases where a parameter is positive- (negative-) definite, NCI automatically works with the natural logarithm of the variable (absolute value of the variable). This practice guarantees that the density estimate is zero for negative (positive) values of the parameter, as it should be. The result is typically a slight improvement in the evaluation metrics.

NCI provides unbiased estimates of the table entries using cross-validation (Hills 1966; Leka & Barnes 2003b; Leka et al. 2018); previously NCI relied upon "n-1" method but now performs cross-validation based on active-region number. For the results here, the last digit of the AARP number is used to define 10 groups, with which 10-fold cross-validation is performed. This approach is invoked in recognition that for any given AARP, some parameters may not evolve significantly over a day or longer. The goal then of AARP-based cross-validation is to avoid using samples of the same AARP to both construct the probability density functions and then use them to predict a sample from the same AARP.

### 3.2.3. Evaluation Metrics

The classifications made by NCI are evaluated using standard quantitative metrics (Jolliffe & Stephenson 2012), to answer the question, "how well did the classifier separate the samples drawn from the two known populations?" NCI reports a large selection of metrics and graphical tools for interpretation; here we focus on a few that are most informative for the present study.

The native results from NCI are the probabilities for each data point of belonging to one or the other population, hence evaluation metrics based on probabilities are most appropriate. The Brier

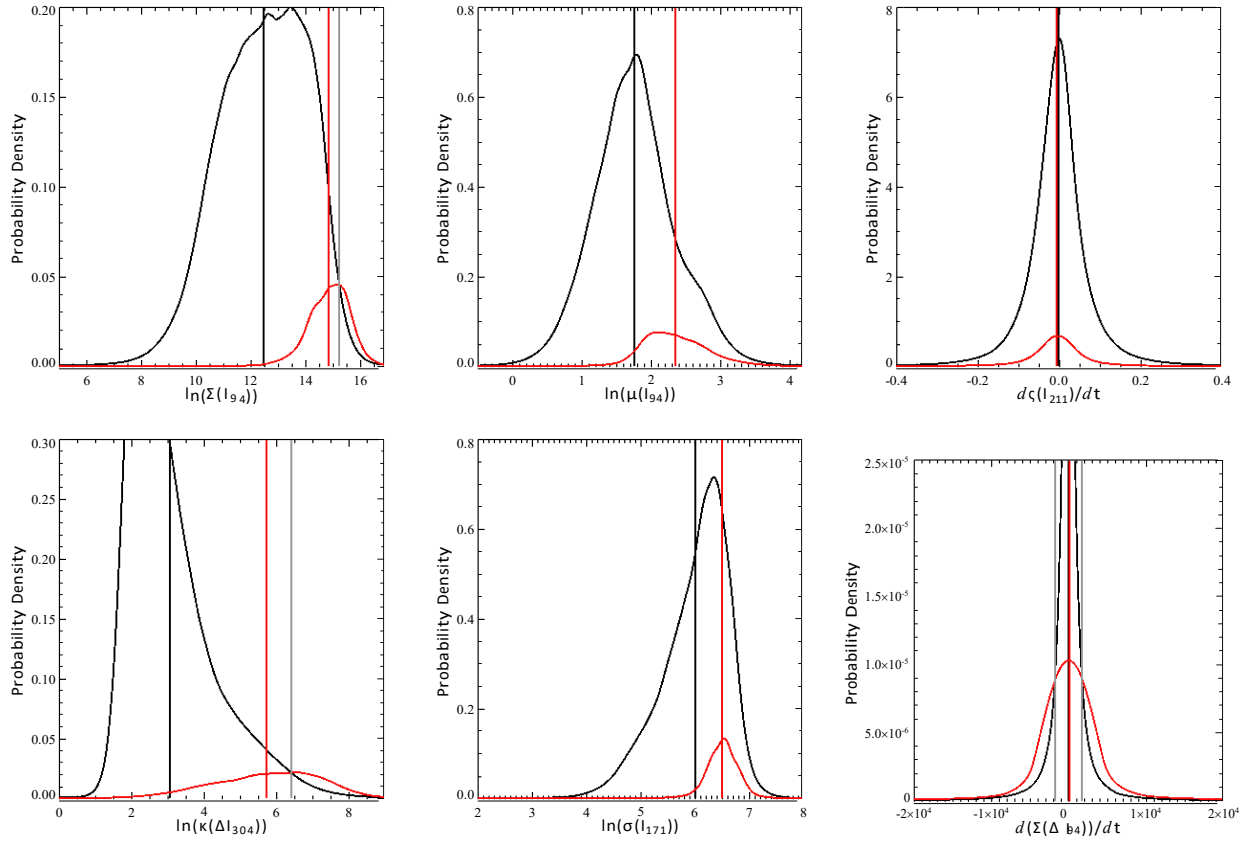



Figure 4. NonParametric Probability Density Functions of six parameters for the C1.0+/24 hr event definition: (Top, left-to-right): the natural log of the total of the 94A˚ emission ($\ln \Sigma(I_{94})$), the natural log of the mean of the 94A ˚emission ($\ln \mu(I_{94})$), and the change with time of the skew of the 211˚A emission ($d\varsigma(I_{211})/dt$). (Bottom, left-to-right): the natural log of the kurtosis of the running-difference of 304˚A images ($\ln \kappa(\Delta I_{304})$), the natural log of the standard deviation of the 171A ˚emission ($\ln \sigma(I_{171})$, c.f. Figure 2), and the change with time of the total of the 94A ˚running-difference images ($d\Sigma(\Delta I_{94})/dt$). For all, event, non-event non-parametric density estimates are shown, their means (- - -/- - -), and the discriminant boundary(ies) which may not be present within the range shown (which itself always encompasses all but the most extreme outliers, if any). See text for discussion.

skill score (B S S) quantifies the performance by normalizing the mean square error of the probability that a point belongs to its true population by the mean square error for the probability based on the "climatology", or ratios of the two population sizes to the total sample size. It is normalized so that "perfect" is 1.0, no skill against the reference is 0.0, and can be negative. B S S effectively summarizes the Reliability Plot ("attributes diagram") that is conditioned on the forecast (classification), and by which sharpness and resolution can be judged; we report the B S S and present Reliability plots in Section 4.

With the assignment of a Probability Threshold ($P_{thr}$) above/below which the resulting probability is deemed to belong to one or the other population, categorical metrics are available (see the discussions in Barnes et al. 2016; Leka et al. 2019a). For these, a classification table is first constructed according to the assigned probability that a data point belongs to one or the other populations, given an assigned $P_{thr}$). Four entries (for 2-option classification) then comprise the classification

table: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). As we are not providing any kind of custom forecasts, we use $P_{thr} = 0.5$ by default, which maximizes the number of correct classifications when the prior probabilities are set proportional to the sample sizes, and is appropriate for physics-interpretable research.

The popular True Skill Statistic (TSS), also known as the Peirce Skill Score (PSS) or Hanssen & Kuiper Skill Statistics (H&KSS) (see Bloomfield et al. 2012; Barnes et al. 2016; Leka et al. 2019a, for discussions) is the difference between the probability of detection (hit rate) and the probability of false detection (false alarm rate). As with all skill scores, it is normalized such that for perfect differentiation TSS= 1.0, while no power to discriminate the populations produces TSS= 0.0. Changing the sample sizes does not impact TSS provided the samples have been drawn from the same populations. "Optimal TSS" or "Maximum TSS" scores are often reported, and are generally earned by setting $P_{thr} \approx$ the event rate R (Table 2) where $R = n_{TP} + n_{FN}/N$ and N is the sample size (Bloomfield et al. 2012; Barnes et al. 2016; Kubo 2019). We report here Max(TSS) with $P_{thr} = R$.

Finally, by calculating the hit rate (POD) and false alarm rate (POFD), the two components of the TSS, through the range of $P_{thr}$ one builds a Receiver (Relative) Operating Characteristic Curve (ROC) plot (see examples and discussion in Leka et al. 2019a). The ROC plot illustrates the ability of a forecast (or classification) to differentiate between events and non-events, and is observation-conditioned. This plot is then summarized by the ROC Skill Score (ROCSS; or the Gini Coefficient) that is related to the ROC area or popular Area Under the Curve (AUC) metric: $G = 2 \cdot AUC - 1.0$ where $G = 1.0$ denotes a perfect score and $G < 0$ indicates worse than zero-skill performance. We report here G but also present ROC plots for a few examples.

We sort the parameters based on the BSS metric, the only metric for which we perform 100-draw bootstrap with replacement (Efron & Gong 1983; Jolliffe & Stephenson 2012; Leka et al. 2018), also based on the last digit of the AARP number, to provide an estimate of the uncertainty in the metric. That is, for each draw independently, the probability density estimates for each population are calculated (Figure 4) and used to generate a probability of an event occurring. This probability varies (usually only slightly) between the different draws, leading to a range of values for the BSS, and slightly moving the location of the discriminant boundary, sometimes leading to different classification tables. The standard deviation of the BSS values is used as an estimate of the uncertainty. The other metrics are calculated directly from the probabilities for each data point, computed using cross-validation but no bootstrap. The rank order of the different metrics does not follow identically, but is generally close (Tables 3–6). Our previous investigation on photospheric magnetic field parameters (Leka et al. 2018) found that, for a given event definition and parameter, the uncertainty across a range of skill scores was relatively constant. Thus, the uncertainties quoted for the BSS are likely to be a reasonable representation of the uncertainty in the Max(TSS) and ROCSS.

### 3.3. *Sample Size and Statistical Flukes*

With the large number of parameters being considered, it is possible that a few parameters may falsely appear to be successful at classifying the data solely by happenstance of this particular sample. The likelihood of this happening is diminished with large sample sizes, but for the M1.0+/24 hr and especially the M1.0+/6 hr event definitions, it may become a concern.

In Barnes et al. (2014), a Monte Carlo experiment was described that draws two random samples from the same population with sizes equal to the sample sizes in question (*e.g.* of the event and non-event samples). In the experiment, the same analysis is performed as on the actual parameters for

fifty times as many parameters as were in the investigation, to more accurately capture the range of possible outcomes. The experiment was performed where the population was a normal distribution, a Cauchy distribution, and a cosine distribution. The resulting distributions of skill scores for the experiment, where no difference is expected, were then compared to the distribution found for the real experiment, and the probability of finding outliers was estimated. In other words, this approach determines the number of statistical outliers that may be expected were there *no* difference in the two underlying populations.

When this experiment was applied to the AARP-matched HARP-based magnetic field parameters in a similar context as the present study and with a similar sample size (but indeed for a larger number of parameters than is being tested here Leka et al. 2018), we estimated there would be <1% chance of a resulting BSS > 0.001/0.002/0.003 by chance alone for single variable NPDA for C1.0+/M1.0+/X1.0+ flares, respectively. Hence we are confident that the results shown here are not particularly susceptible to statistical flukes.

Additionally, the bootstrap provides an uncertainty for the BSS. As discussed in Section 4.2, for the top performing results and indeed for most parameters across the C1.0+ and M1.0+/24 hr event lists, the reported BSS are at the 5σ, 10σ or higher detection level. For M1.0+/6 hr which is the experiment with the smallest "yes-event" sample size and the smallest event rate, the BSS scores are smaller, barely above 0.0, although the bootstrap-derived uncertainties are only a factor of 2 larger (see Section 4.2). Even with almost a solar-cycle's worth of data, the sample of larger events that occur within 6 hours of any given time of day is, statistically speaking, very small.

## 4. RESULTS

In these sections we highlight some examples and call out the best and the worst performing parameters in order to give an overview of the results. All computed parameters, and resulting probabilities are available (Leka et al. 2022), so readers can examine the distributions for other parameters of interest, and (for example) compute additional skill scores or apply other analysis methods to the data.

### 4.1. *NonParametric Density Estimates*

We show in Figure 4 the nonparametric density estimates for a selection of parameters, all for the C1.0+/24 hr definition primarily because the distributions of both populations are clearly visible; the class imbalance between events and non-events for the other definitions (Table 2) simply make presentation more challenging.

There is quite a range of distribution shapes amongst the parameters. For one of the most intuitive parameters, $\Sigma(I_{94})$ (Figure 4 top left), the density estimates are distinctly offset from each other, and there is a single discriminant boundary to the right of which the events have a higher probability than the non-events. In the next two parameters $\mu(I_{94})$ and $d\varsigma(I_{211})/dt$ (Figure 4 top middle and right, respectively) there is no discriminant boundary; for the former, even though the distributions are distinctively offset from each other (the means are visibly different), the low event rate (large class imbalance) means that the event probability never exceeds the non-event probability whereas in the latter, there is almost no difference in the event *vs.* non-event distribution means or shapes. Despite the lack of a discriminant boundary, $\mu(I_{94})$ still has significant skill as measured by the BSS (BSS= $0.084 \pm 0.006$), while $d\varsigma(I_{211})/dt$ does not.

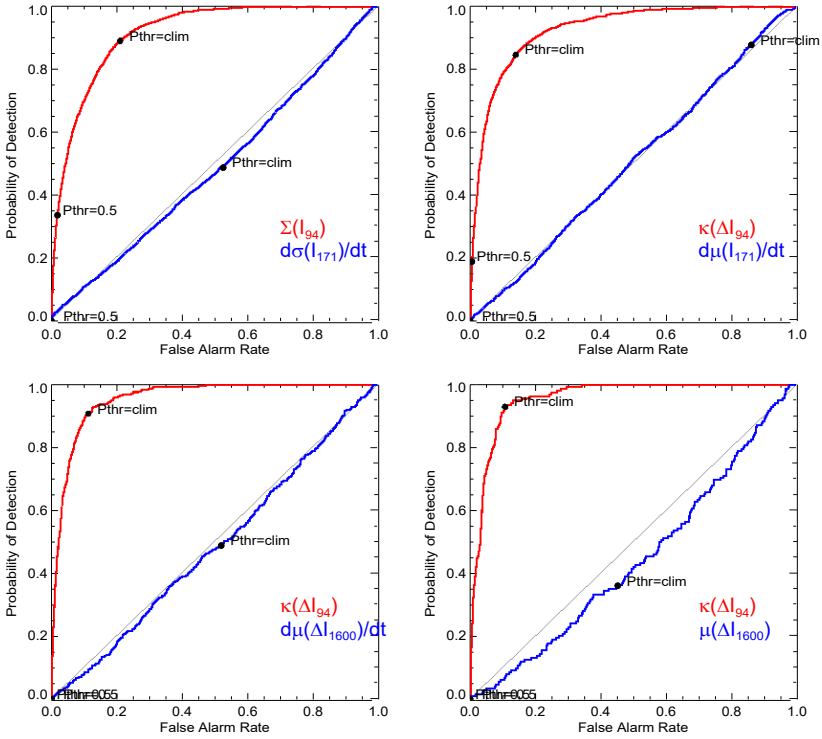



Figure 5. Receiver Operating Characteristic (ROC) plots for (Top, Left/Right) C1.0+/24 hr, C1.0+/6 hr (Bottom Left/Right) M1.0+/24 hr, M1.0+/6 hr, for the parameters as indicated. For all, the 'climatology' probability threshold is as listed in Table 2. The parameters shown are generally the top- and bottom-scoring parameters by G, c.f. Tables 3–6.

The first two parameters in the bottom row of Figure 4 show similar behavior to the corresponding parameters in the top row: $\kappa(\Delta I_{304})$ provides a single clear discriminant boundary and very different distributions, while the distributions for the $\sigma(I_{171})$ samples are reminiscent of the $\mu(I_{94})$ distributions, again the population distributions are distinguishable (the means are well separated), there is significant skill, but there is no discriminant boundary. Finally, the $d\Sigma(\Delta I_{94})/dt$ distributions are centered exactly the same, however, unlike $d\varsigma(I_{211})/dt$, there are two discriminant boundaries because the event population is wider than the non-event population.

### 4.2. *Metrics Scores and Evaluation Plots for AARP-based Parameters*

The results are sorted on BSS, and we present the top-10 and bottom-5 BSS-scoring parameters in Tables 3 – 6; the full results are available in machine-readable format. For each of the parameters we also compute the "Max(TSS)" (with $P_{thr} = R$) and the ROCSS or G. The order of the parameters based on the latter scores does not exactly follow the ordering of the BSS, but does so loosely, especially considering the bootstrap-based uncertainties for the BSS.

While the BSS and G summarize the Reliability and ROC plots respectively, it is instructive to see the behaviors explicitly by which to judge bias, *etc*. ROC plots (Figure 5) and Reliability plots (Figure 6) are shown for one of the best and one of the worst-scoring parameters each (according to BSS, as per Tables 3 – 6), for each event definition.

Overall, the classification results for select UV/EUV parameters show confidence at statistically significant levels for the C1.0+/24 hr, C1.0+/6 hr, and M1.0+/24 hr event definitions. By this we

Table 3. Results: C1.0+/24 hr

| Top 10 Scoring Parameters: C1.0+/24 hr | | | |
|---|---|---|---|
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $\kappa(\Delta I_{94})$ | 0.332 ± 0.011 | 0.650 | 0.816 |
| $\kappa(\Delta I_{131})$ | 0.315 ± 0.011 | 0.658 | 0.810 |
| $\kappa(\Delta I_{171})$ | 0.312 ± 0.012 | 0.670 | 0.809 |
| $\kappa(\Delta I_{304})$ | 0.310 ± 0.010 | 0.668 | 0.812 |
| $\Sigma(I_{94})$ | 0.302 ± 0.013 | 0.680* | 0.830* |
| $\kappa(\Delta I_{193})$ | 0.301 ± 0.011 | 0.657 | 0.814 |
| $\kappa(\Delta I_{211})$ | 0.291 ± 0.011 | 0.651 | 0.794 |
| $\Sigma(\Delta I_{94})$ | 0.286 ± 0.011 | 0.626 | 0.788 |
| $\Sigma(I_{335})$ | 0.280 ± 0.014 | 0.672 | 0.822 |
| $d\Sigma(\Delta I_{94})/dt$ | 0.273 ± 0.011 | 0.597 | 0.761 |
| Bottom 5 Scoring Parameters: C1.0+/24 hr | | | |
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $d\mu(I_{193})/dt$ | 0.001 ± 0.000 | 0.035 | 0.046 |
| $\mu(\Delta I_{171})$ | 0.001 ± 0.001 | 0.052 | 0.047 |
| $d\sigma(I_{171})/dt$ | 0.000 ± 0.000 | 0.023 | 0.015 |
| $d\kappa(I_{131})/dt$ | −0.011 ± 0.009 | 0.203 | 0.283 |
| $d\kappa(I_{335})/dt$ | −0.068 ± 0.028 | 0.160 | 0.262 |

*: Top or Bottom score for Max(TSS) and for G. In this case the worst Max(TSS)= -0.038, and G= -0.034 both for $d\varsigma(I_{211})/dt$ which has BSS= 0.001 ± 0.001.

Note—Table 3 is published in its entirety in machine-readable format. A portion is shown here for guidance regarding its form and content.

mean that the sample sizes are large enough that the bootstrap-derived uncertainties in the BSS, plus the AARP-focused cross validation, provide good estimates of the uncertainties and that the BSS results indicate skill above climatology (BSS> 0.0). We did not perform a separate bootstrap or sorting for the other metrics provided, but assume that the (un)certainty levels are similar. As has been found in other studies, there are numerous parameters that perform similarly within the error bars.

The uncertainties related to the BSS results are overall small especially compared to the BSS results for C1.0+/24 hr and C1.0+/6 hr. For M1.0+/6 hr, while the larger error bars reflect a smaller sample of events, the BSS results barely indicate skill above the climatology. The reliability plots (Figure 6) for the better performing metrics do show a good correspondence between the predicted probabilities and the observed frequency of occurrence, the points generally falling within their $1\sigma$ error bars of the x = y line. In other words, even for the M1.0+/6 hr events and even with their low BSS, the predictions are “reliable”. However, the vast majority of the predictions (especially for the M1.0+/24 hr and M1.0+/6 hr events) are probabilities close to the event rates, and this lack of sharpness is reflected in the low BSS.

However, the G results are quite high, generally, as are the Max(TSS). For rare events, as displayed in the ROC plots (Figure 5), the metrics reward a high probability of detection at the expense of an

Table 4. Results: M1.0+/24 hr

| Top 10 Scoring Parameters: M1.0+/24 hr | | | |
|---|---|---|---|
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $\kappa(\Delta I_{94})$ | 0.160 ± 0.015 | 0.794* | 0.909* |
| $\kappa(\Delta I_{131})$ | 0.132 ± 0.010 | 0.734 | 0.862 |
| $\Sigma(\Delta I_{94})$ | 0.131 ± 0.015 | 0.704 | 0.840 |
| $d\kappa(\Delta I_{94})/dt$ | 0.125 ± 0.018 | 0.680 | 0.837 |
| $\Sigma(\Delta I_{131})$ | 0.118 ± 0.021 | 0.640 | 0.786 |
| $\kappa(\Delta I_{211})$ | 0.117 ± 0.008 | 0.750 | 0.863 |
| $d\varsigma(\Delta I_{94})/dt$ | 0.116 ± 0.009 | 0.640 | 0.802 |
| $\kappa(\Delta I_{304})$ | 0.116 ± 0.010 | 0.725 | 0.851 |
| $d\varsigma(\Delta I_{131})/dt$ | 0.110 ± 0.016 | 0.658 | 0.812 |
| $d\Sigma(\Delta I_{131})/dt$ | 0.109 ± 0.017 | 0.626 | 0.764 |
| Bottom 5 Scoring Parameters: M1.0+/24 hr | | | |
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $d\sigma(I_{171})/dt$ | 0.000 ± 0.000 | 0.049 | 0.038 |
| $d\mu(I_{171})/dt$ | 0.000 ± 0.000 | 0.028 | 0.019 |
| $d\mu(\Delta I_{1600})/dt$ | 0.000 ± 0.000 | -0.031* | -0.036* |
| $\mu(\Delta I_{1600})$ | 0.000 ± 0.000 | -0.031 | -0.024 |
| $d\mu(I_{131})/dt$ | −0.002 ± 0.004 | 0.157 | 0.197 |

*: Top or Bottom score for Max(TSS) and for G.

Note—Table 4 is published in its entirety in machine-readable format. A portion is shown here for guidance regarding its form and content.

increased false alarm rate. Thus the predictions have good ability to distinguish between the event and non-event populations, or good resolution.

Overall, the class imbalance in all event definitions, but especially the M1.0+/24 hr and M1.0+/6 hr as we define them here, is extreme. This can lead to impressive Max(TSS) scores. Simultaneously, the BSS is negatively impacted by the class imbalance although it takes the climatology into account since the climatology provides the reference prediction.

The best-performing parameters across the four event definitions are dominated by the kurtosis of the running-difference images. The kurtosis detects deviation from a Gaussian distribution in terms of central peak *vs.* wing relative strength. An enhanced kurtosis or leptokurtic distribution, which is associated with an increased probability of flaring, has an over-population of the wings relative to a normal distribution, although it can also indicate an under-population of the central peak (and *vice versa* for a low kurtosis or platykurtic distribution). In terms of moments, the remaining best-performing parameters are typically either the skew or the total of the running-difference images.

There are fewer direct-image (*vs.* running-difference image) and evolution ("dX/dt") parameters than expected in the top-10 across event definitions (fewer than 5 of 10); evolution-based parameters in fact tend to dominate the low-scoring BSS results. As mentioned in Section 3.2.3, the "dX/dt" parameters may be more susceptible to outliers, and looking beyond the top-10 their frequency becomes higher although running-difference images still dominate over direct images. The cos(θ)

Table 5. Results: C1.0+/6 hr

| Top 10 Scoring Parameters: C1.0+/6 hr | | | |
|---|---|---|---|
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $\kappa(\Delta I_{94})$ | 0.247 ± 0.010 | 0.703* | 0.853* |
| $\kappa(\Delta I_{131})$ | 0.214 ± 0.010 | 0.684 | 0.828 |
| $\Sigma(\Delta I_{94})$ | 0.207 ± 0.012 | 0.675 | 0.816 |
| $\kappa(\Delta I_{211})$ | 0.203 ± 0.008 | 0.669 | 0.818 |
| $\kappa(\Delta I_{171})$ | 0.199 ± 0.008 | 0.685 | 0.828 |
| $\kappa(\Delta I_{304})$ | 0.199 ± 0.007 | 0.681 | 0.820 |
| $\kappa(\Delta I_{193})$ | 0.199 ± 0.008 | 0.676 | 0.829 |
| $d\Sigma(\Delta I_{94})/dt$ | 0.196 ± 0.014 | 0.622 | 0.775 |
| $\varsigma(\Delta I_{94})$ | 0.192 ± 0.012 | 0.567 | 0.717 |
| $d\varsigma(\Delta I_{94})/dt$ | 0.184 ± 0.012 | 0.577 | 0.735 |
| Bottom 5 Scoring Parameters: C1.0+/6 hr | | | |
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $d\mu(I_{193})/dt$ | 0.000 ± 0.000 | 0.027 | 0.050 |
| $d\mu(I_{171})/dt$ | 0.000 ± 0.000 | 0.014* | 0.002* |
| $d\sigma(I_{171})/dt$ | 0.000 ± 0.000 | 0.021 | 0.027 |
| $d\kappa(I_{211})/dt$ | 0.000 ± 0.004 | 0.217 | 0.269 |
| $d\kappa(I_{335})/dt$ | −0.066 ± 0.024 | 0.188 | 0.320 |

*: Top or Bottom score for Max(TSS) and for G.

Note—Table 5 is published in its entirety in machine-readable format. A portion is shown here for guidance regarding its form and content.

location (observing angle) parameter shows minimal but not zero classification power. This result is due to the HARP selection criteria that includes numerous small plage regions at greater absolute latitudes than spot-containing active regions. These plage regions generally belong to the "no-event" population, providing a small discriminating advantage to the middle latitudes and the corresponding cos(θ) ranges.

#### 4.2.1. Wavelength-compared Classification Performance

The different filters of AIA are sensitive to plasma at different temperatures, and often sensitive to more than one temperature (Lemen et al. 2012). The behavior of the plasma in the corresponding physical regimes may reflect different thermal or density responses to energy build up, or different kinematic responses to photospheric driving motions, for example. To address these questions, we first simply evaluate the parameters' performance as grouped by wavelength; in Section 5.3 we discuss more the physical implications of the results.

A cursory look at Tables 3–6 gives the impression that filters which detect hotter plasma more frequently appear in the "Top-10", across event definitions. The C IV 1600A-based parameters are never in the "top-10", the He II 304A- and Fe IX 171A-based parameters do make the top tiers in BSS but rarely. The top parameters are dominated by parameters built from the Fe XVIII 94A filter and the other filters sensitive to hotter plasma, for example the Fe XXI-sensitive 131A filter.

Table 6. Results: M1.0+/6 hr

| Top 10 Scoring Parameters: M1.0+/6 hr | | | |
|---|---|---|---|
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $\kappa(\Delta I_{94})$ | 0.070 ± 0.014 | 0.821* | 0.913* |
| $\sigma(I_{94})$ | 0.070 ± 0.018 | 0.707 | 0.860 |
| $d\varsigma(\Delta I_{131})/dt$ | 0.067 ± 0.017 | 0.701 | 0.846 |
| $\Sigma(\Delta I_{131})$ | 0.061 ± 0.023 | 0.646 | 0.806 |
| $\varsigma(\Delta I_{131})$ | 0.058 ± 0.014 | 0.720 | 0.819 |
| $d\Sigma(\Delta I_{131})/dt$ | 0.056 ± 0.011 | 0.624 | 0.810 |
| $\kappa(\Delta I_{131})$ | 0.056 ± 0.011 | 0.778 | 0.886 |
| $\Sigma(\Delta I_{94})$ | 0.055 ± 0.017 | 0.708 | 0.836 |
| $\sigma(\Delta I_{131})$ | 0.054 ± 0.019 | 0.575 | 0.757 |
| $\varsigma(\Delta I_{94})$ | 0.054 ± 0.019 | 0.661 | 0.774 |
| Bottom 5 Scoring Parameters: M1.0+/6 hr | | | |
| Parameter | Brier Skill Score | Max(TSS) | G or ROCSS |
| $d\sigma(I_{171})/dt$ | 0.000 ± 0.000 | 0.042 | 0.036 |
| $\mu(I_{171})$ | 0.001 ± 0.002 | 0.268 | 0.340 |
| $\mu(\Delta I_{1600})$ | 0.001 ± 0.001 | 0.060 | 0.046 |
| $\mu(I_{1600})$ | −0.001 ± 0.001 | 0.141 | 0.139 |
| $d\mu(I_{131})/dt$ | −0.003 ± 0.006 | 0.150 | 0.259 |

*: Top score for Max(TSS) and for G. In this case the worst Max(TSS)= -0.104, G = −0.116 both for $d\mu(I_{171})/dt$ which has BSS= 0.000 ± 0.000.

Note—Table 6 is published in its entirety in machine-readable format. A portion is shown here for guidance regarding its form and content.

We note that the top-performing parameters for the C1.0+ event definitions include parameters across all analyzed EUV filters, while for the M1.0+ event definitions the top-ranked parameters are predominantly those derived from 94 and 131 Å filters (Tables 3–6).

In Figures 7, 8, 9, 10 we group the BSS results by wavelength. What is striking in these plots with regards to the performance by different AIA filters is that the 94Å parameters by and large perform consistently well (comparatively speaking), with all "radar sectors" filled in at least somewhat. In contrast, the radar plots for 211Å, for example, have definite gaps; for example, while the $\kappa(\Delta I_{211})$ scores well, the $d\kappa(I_{211})/dt$ parameter does not.

Overall, this presentation confirms the highlights of Tables 3–6: the performance is overall lower for the shorter-validity definitions, and uncertainties are larger for the event definitions that have smaller event-population sample sizes (higher class imbalance). There are more parameters that perform with higher classification success for the 94 Å filter than most of the others, but then the 304 Å parameters also have a fairly high frequency of similar performance (albeit not "high performing" by this metric *per se*). The other AIA filters show a more mixed performance, with the 1600 Å arguably the lowest overall. Notably, for all wavelengths, the kurtosis- and skew- and total-based evaluation of running-difference images are often the highest performing parameters of any particular wavelength.

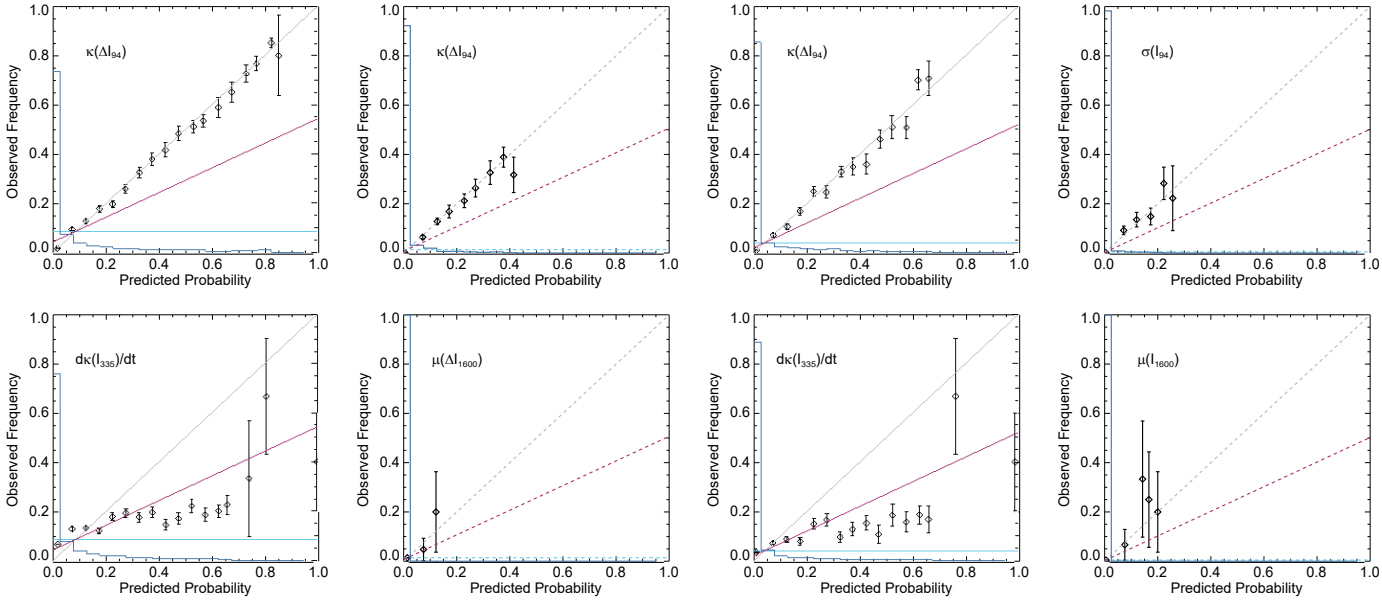



Figure 6. Reliability plots for (Left to Right) C1.0+/24 hr, M1.0+/24 hr, C1.0+/6 hr, M1.0+/6 hr for top-performing parameters (Top) and low-performing parameters (Bottom), according to BSS, as indicated. The x = y line indicates perfect reliability, the histogram (blue) is the frequency of occurrence for each prediction bin, the horizontal line (light-blue dashed) indicates the climatology (no resolution) and the "no skill" line is also plotted (red dashes). The 1σ error bars are shown, and reflect the sample size in each bin.

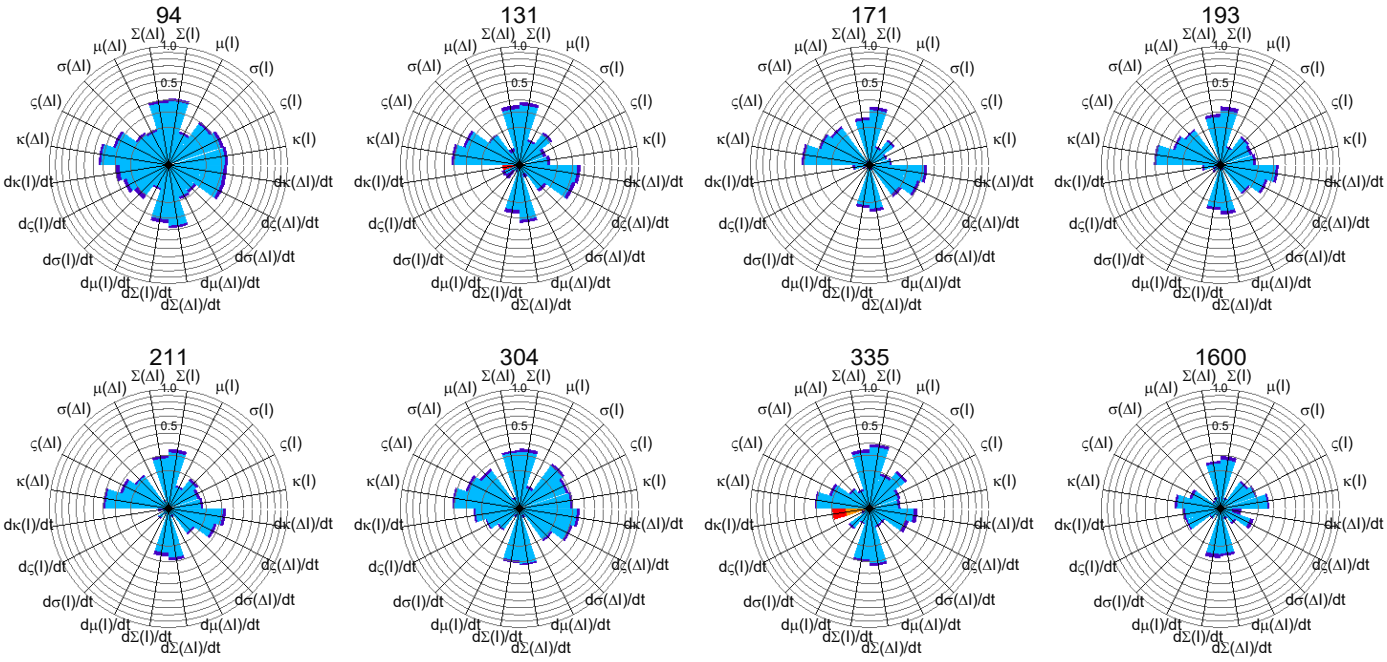



Figure 7. Radar plots showing Brier Skill Score for all parameters, grouped by filter, as labeled. Arcs indicate the range of the $BSS \pm \sigma_{BSS}$, BSS> 0 (blue) and |BSS| for BSS< 0 (orange), with darker hues indicating the uncertainty ranges. Shown: C1.0+/24 hr event definition results.

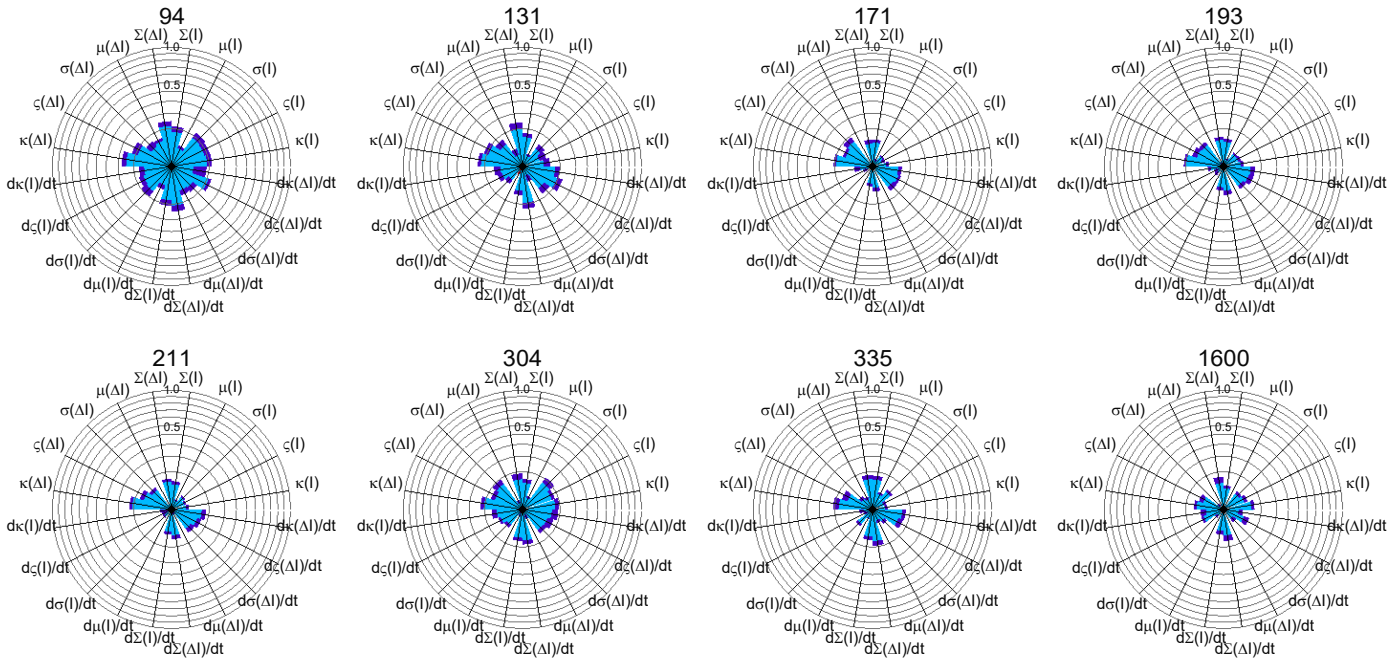



Figure 8. Same as Figure 7 for the M1.0+/24 hr event definition results.

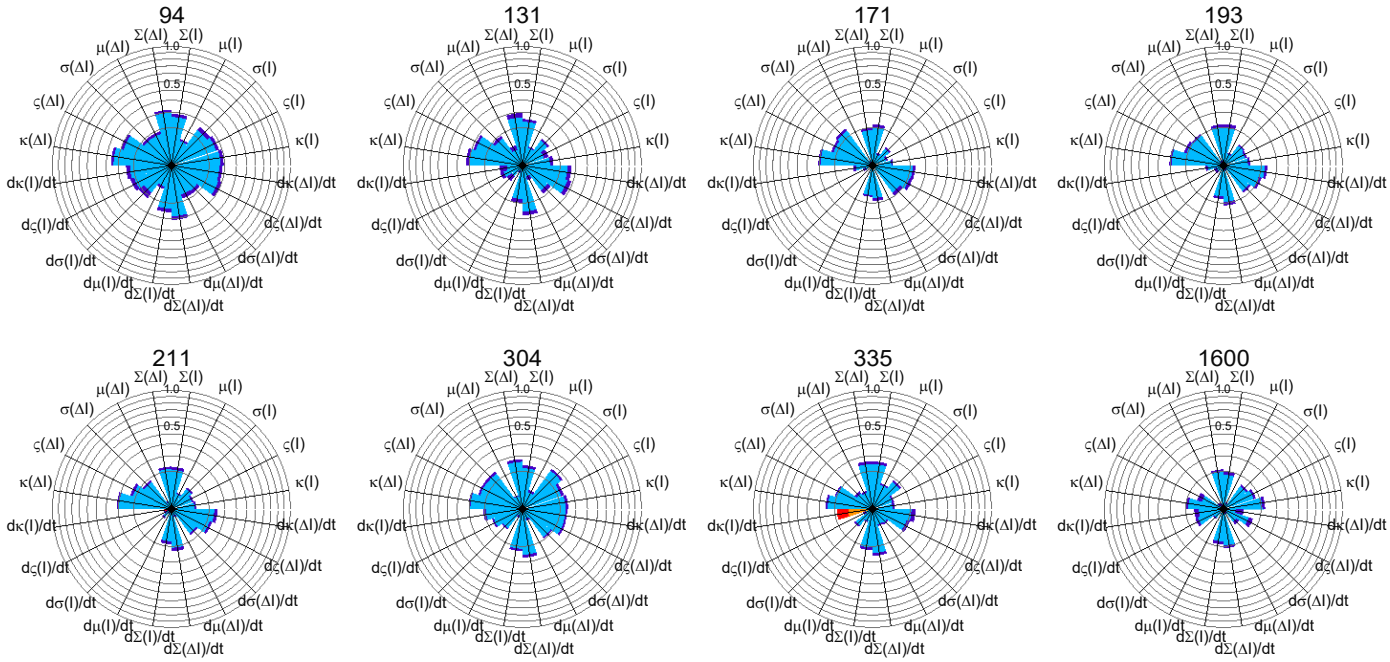



Figure 9. Same as Figure 7 for the C1.0+/6 hr event definition results.

#### 4.2.2. Performance Changes between Event Definitions

Generally speaking, the BSS scores decrease while Max(TSS) and G stay the same or increase between C1.0+ and M1.0+ definitions, and between, for example, the 24 hr and 6 hr validity times. This is fairly evident as a general rule from the discussion thus far and is not unexpected given the

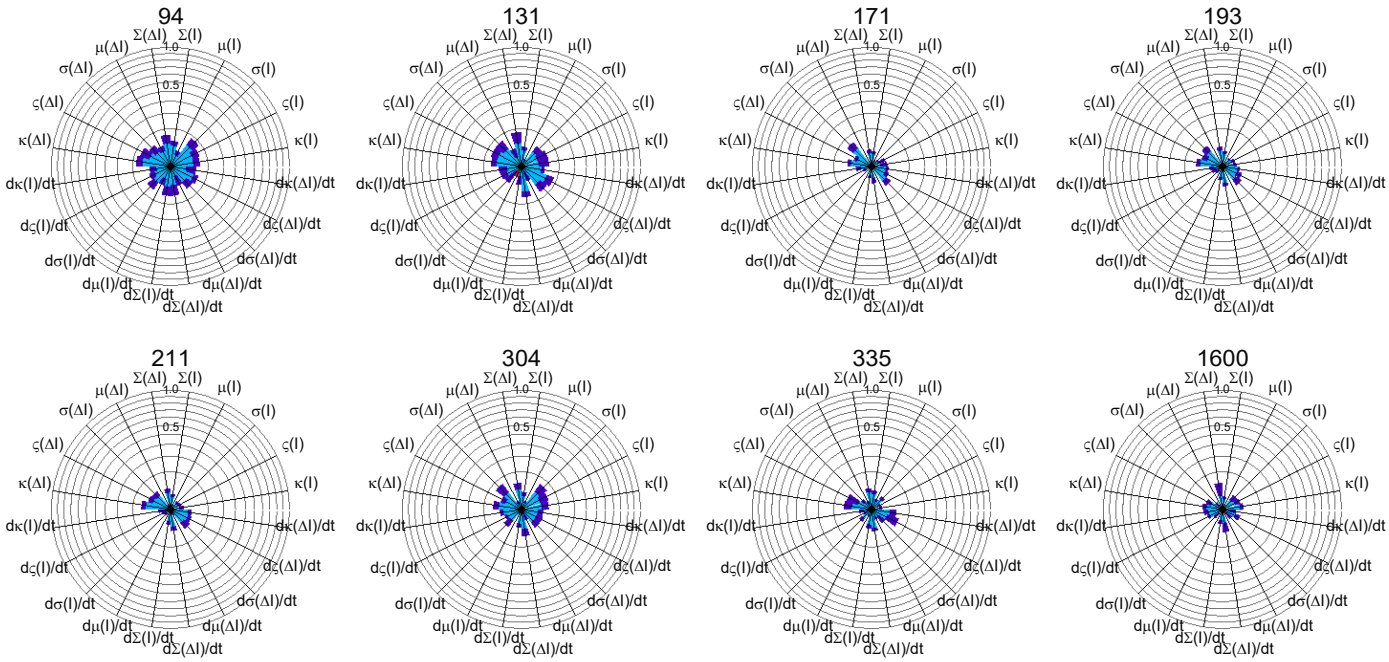



Figure 10. Same as Figure 7 for the M1.0+/6 hr event definition results.

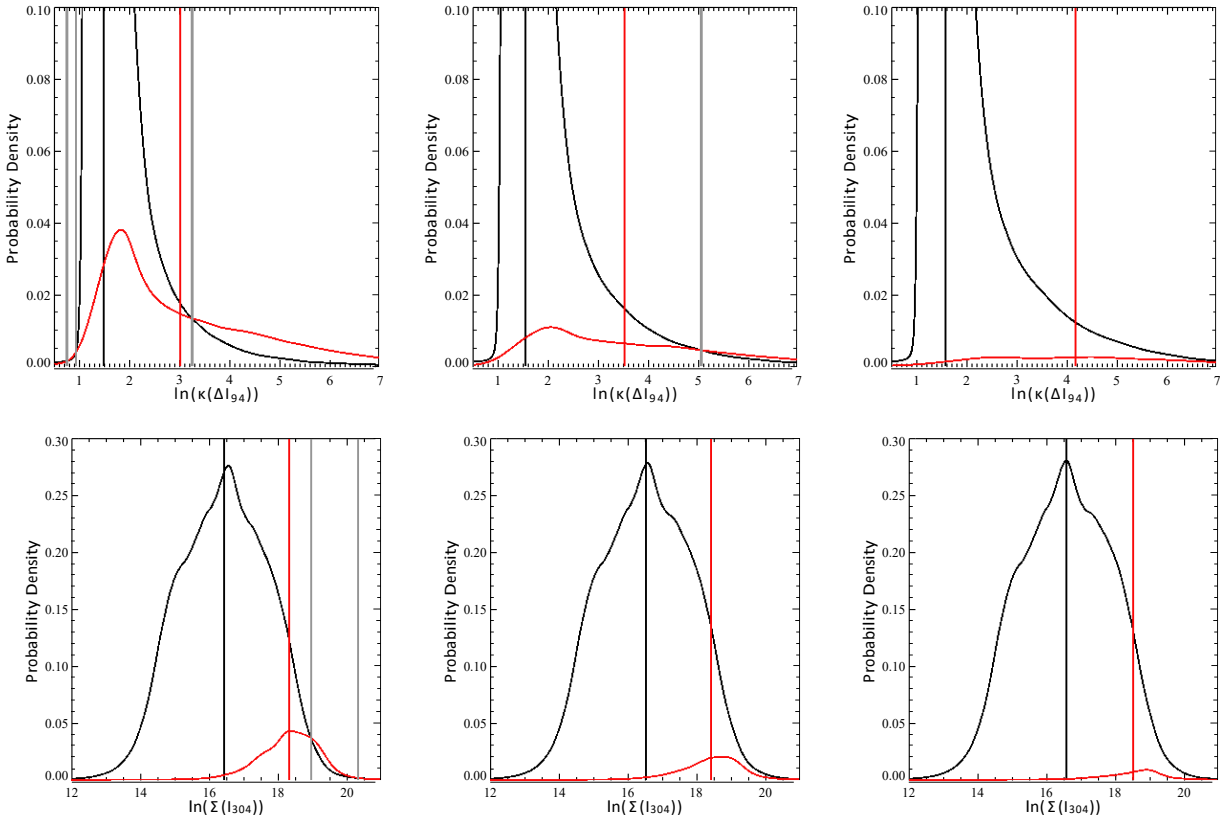
sensitivity of B S S to event rates and relative insensitivity of Max(TSS) to the same (Bloomfield et al. 2012; Barnes et al. 2016).

However, there are some variations in this behavior. There are some parameters for which the relative distributions (event- *vs.* non-event) vary noticeably with an expected increase in Max(TSS) between, for example, C1.0+ and M1.0+ definitions - reflecting a shift to higher parameter values for the event population, for example (Figure 11, top), and a relatively smaller decrease in the B S S. For other parameters, the distributions vary in relative magnitude reflecting the different relative sample sizes, but the distribution means, for example, do not significantly change (Figure 11, bottom). In this case, the Max(TSS) does not appreciably change because the change in magnitude is offset by the change in the value of R, and the value of the B S S decreases more substantially. We found no obvious or systematic behavior in this regard between parameter "classes" (those based on direct *vs.* running-difference images, or static *vs.* dX/dt parameters) except that similar parameters often (but not always) behave the same across wavelengths.

### 4.3. *Performance Changes with Solar Cycle*

Solar-cycle-related variations may impact the ability of the parameters generated here to classify flare-imminent active regions. The background UV- and EUV emission (Argiroffi et al. 2008; Schonfeld et al. 2017) may add a constant to the mean or summation-based parameters, and varying event rates can change the prior probabilities (McCloskey et al. 2018; Leka et al. 2019a). Even running-difference images may be subject to subtle changes in signal-to-noise ratios due to high background contamination, potentially impacting their ability to detect changes in active-region structure.

To examine the behavior of these parameters against cycle-related influences, we break the data set into two subsets, first with years that were "active" parts of the cycle (2011-2015 inclusive, plus 2017) and "quiet" (the rest), based partly on the start of the high-activity time as defined by coronal




| Parameter | Event Definition | BSS | Max(TSS) | G or ROCSS |
|---|---|---|---|---|
| $\kappa(\Delta I_{94})$ | C1.0+/24 hr | 0.332 ± 0.011 | 0.650 | 0.816 |
| | C1.0+/6 hr | 0.247 ± 0.010 | 0.703 | 0.853 |
| | M1.0+/24 hr | 0.160 ± 0.015 | 0.794 | 0.909 |
| $\Sigma(I_{304})$ | C1.0+/24 hr | 0.240 ± 0.014 | 0.613 | 0.775 |
| | C1.0+/6 hr | 0.167 ± 0.008 | 0.589 | 0.747 |
| | M1.0+/24 hr | 0.095 ± 0.015 | 0.616 | 0.766 |

Figure 11. NonParametric Probability Density Functions of two parameters (top) $\kappa(\Delta I_{94})$ and (bottom) $\Sigma(I_{304})$ across three event definitions: C1.0+/24 hr, C1.0+/6 hr, M1.0+/24 hr. Presentation is the same as Figure 4. Also shown are the relevant entries for the performance metrics. The distribution of the event and non-event density estimations vary significantly for $\kappa(\Delta I_{94})$ across event definitions, most easily seen by the increase in the mean for the event population, and is reflected in changes in their relative BSS, but for $\Sigma(I_{304})$ the distributions change primarily in amplitude, due to the different prior probabilities from the different event rates, so the differences in performance in particular for Max(TSS) are much less.

temperature (Schonfeld et al. 2017), and partly due to flaring rates. This partitioning provided total sample size of 4898 AARPS (quiet) and 27169 AARPS (active). We run the full analysis, then look in detail for two very different but originally high-scoring parameters, $\Sigma(I_{94})$ and $\kappa(\Delta I_{94})$.

The resulting probability density functions for the quiet and active periods for the C1.0+/24 hr event definition are shown in Figure 12, using equal prior probabilities for clarity. Overall, we find very little difference in the distributions between the subsets. There is a very small shift toward

higher values for the $\Sigma(I_{94})$ parameter during the "active" years, but it shifts for both event- and non-event distributions. There is almost no discernible difference in the distributions for $\kappa(\Delta I_{94})$.

The event rates differ significantly between the active and quiet periods, as designed. The sample sizes under this division are very small in most cases, leading to the situation that the adaptive-kernal NPDA is no longer an appropriate model to use. The quiet period is most susceptible, with the number of events for these years being: M1.0+/24 hr: 15; C1.0+/6 hr: 65; M1.0+/6 hr: 5. These small numbers mean that for those event definitions, we cannot compare the results for active years to those for quiet years with confidence; hence we concentrate on C1.0+/24 hr for the statistical analysis.

In Figure 12 we show scatter plots of the BSS and Max(TSS) for C1.0+/24 hr for the quiet and active periods separately against those scores resulting from the full dataset. For the active subset, the difference against the full dataset is minimal for both metrics. For the quiet subset however, the BSS shows a strong systematic decrease whereas the Max(TSS) shows scatter that is, within the expected uncertainties, without significant trend. Recalling that BSS is sensitive to climatological event rates whereas Max(TSS) is not (Jolliffe & Stephenson 2012), we demonstrate that the varying event rates have a measureable impact on some evaluation metrics.

Combining this result with the minimal differences in the probability densities between quiet and active parts of the solar cycle, we conclude that cycle-related event-rate variations have a much larger impact on the ability to classify our parametrizations, as measured by some metrics, than the impact of variation in background emission.

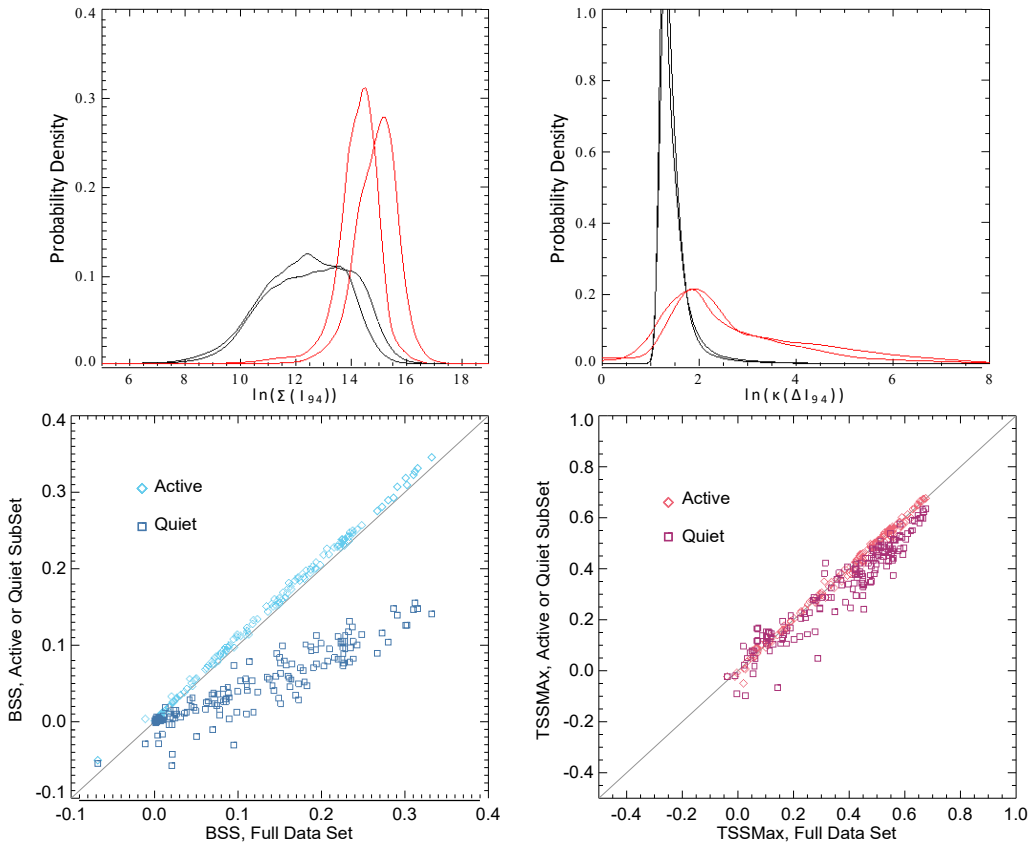
## 5. INTERPRETATION

Because we construct the parametrizations ourselves, they enable physical interpretation to the extent allowed by analysis of just the images themselves. The span of regimes sampled, in temperature/density, height, and temporal dimensions, provides the opportunity to understand the causes and effects of upper-atmosphere behavior in this context.

### 5.1. *Temporal Variability*

The parameterizations examine the variability of the corona on two different time scales. All of the $\Delta I_{\square}$ parameters look at the variation in intensity on 72 s cadence which tracks both small-scale short-lived brightening events and (dis)appearances and kinematics of structures including coronal loops. The moments of the running-difference images $M(\Delta I_{\square})$ further quantify the behavior: increased or decreased mean indicates a preferential brightening or dimming on these timescales, or the appearance / disappearance of structures. The standard deviation indicates the spatial (lack of) quietness. the skew and kurtosis provide sensitivity to the far wings of the distributions indicating small-scale dynamics related to temperature changes or to kinematic variations.

The $M(\Delta I_{\square})$ overwhelmingly dominate the top-10 performing parameters across all event definitions, and in particular the higher-order moments $\varsigma(\Delta I_{\square})$, $\kappa(\Delta I_{\square})$. The density estimates (see example in Figure 4) show enhanced kurtoses for the event populations relative to the non-event populations, indicating wing enhancements rather than degradation of the distribution peaks. Consistently high kurtosis over the 13 min indicates continual presence of rapidly-changing but large-amplitude brightness fluctuations (see Figure 1). In contrast, the $\mu(\Delta I_{\square})$, $\sigma(\Delta I_{\square})$ parameters that should be sensitive to more subtle variations such as expected from gradual loop motion, do not generally perform well in BSS although some have notable G. Non-activity-related intensity changes as due to gradual




| Event Definition | Subset | # Events, (Events Rate R) | median %change, vs. full set, BSS | median %change, vs. full set, Max(TSS) |
|---|---|---|---|---|
| C1.0+/24 hr | Quiet | 178 (0.036) | -61 | -11 |
| | Active | 2574 (0.095) | +6.1 | -0.02 |

Figure 12. [Top]: Probability density functions for C1.0+/24 hr event- and non-event distributions comparing the "quiet" part of the solar cycle (dashed) with the "active" part of the solar cycle (solid). Equal prior probabilities are used to highlight differences in the shape of the PDEs, separate from the changes due to the different event rates. (Left): $\Sigma(I_{94})$, (Right): $\kappa(\Delta I_{94})$. [Bottom]: Comparisons of BSS (left) and Max(TSS) (right) for all parameters, showing results for the quiet- and active- parts of the cycle (as indicated) vs. the metrics for the full data set. The table summarizes the subset characteristics and the resulting differences for two metrics for the C1.0+/24 hr event definition.

loop motion or gradual loop heating / cooling generally proceed slower than the cadence here, and additionally involve preferentially larger (full-loop) structures (Viall & Klimchuk 2012). Hence, there is strong indication, from multiple parameter results, that enhanced variability in brightness or enhanced kinematic activity, on short timescales and small spatial scales, is a discriminating feature of flare-imminent active regions.

Longer-term evolution is reflected in the slope of the linear fit to the 7 hourly samples (Figure 2). We find that, for example, for C1.0+/24 hr the $d\Sigma(\Delta I_{94})/dt$ parameter performs well, and has discriminant boundaries in the wings of the distribution (Figure 4). This means that impending activity is indicated by either a rapid increase or a rapid decrease in the level of rapid intensity fluctuations in the 94A filter. ˚A similar scenario is found for M1.0+/24 hr for $d\varsigma(\Delta I_{94,131})/dt$, the temporal

variation of the skew of the running-difference analysis from 94Å and 131Å filter images (Table 4): on longer timescales, either increasing *or decreasing* levels of short-term brightness variability can indicate upcoming activity (Table 4). In contrast, the $d\varsigma(I_{211})/dt$ parameter for the C1.0+/24 hr definition performs very poorly and in Figure 4 it is easy to see why: the event and non-event density estimates are essentially identical, apart from the different prior probabilities.

Across the event definitions, parameters describing the evolution on hours-long timescales (the dX/dt parameters) are not generally overall better- or worse- performing than the static parameters. They do not appear as frequently as would be expected by even chance in the top-most tiers of performance, but have BSS that are within the uncertainties of many static parameters, and *vice versa*. In other words, while in certain cases for certain parameters and certain event definitions there may be a dX/dt parameter that shows promise for relating coronal evolution to imminent flare activity, there will be at least a few other parameters that do *not* track the evolution but which perform as well. The results here show a small preference for static parameters, but we note this may be a result of outliers rather than a true property of the Sun.

### 5.2. *The Totals, The Moments*

The extensive $\Sigma(I_{\square})$ parameters scale with the size of the AARP, whereas the intensive parameters (the moments $M(I_{\square})$) do not. We see here that extensive parameters can perform at least as well as some of the intensive parameters. In addition to $\Sigma(I_{94})$ being a "top-10" discriminator for C1.0+/24 hr (see also Figure 4), most $\Sigma(I_{\lambda})$ parameters for EUV wavelengths (meaning, all but 1600Å) have high ranking across the event definitions. The general ability of extensive AIA-based parameters to differentiate between flare-imminent and flare-quiet groups is consistent with the results of numerous prior studies, in particular those based on the photospheric magnetic flux (reflecting long-held observers' wisdom) that, simply put, "size matters" (see discussions in Sawyer et al. 1986; Leka & Barnes 2003b, 2007). Larger active regions have more total emission in the corona and chromosphere (as heating functions are believed to scale with magnetic flux, (e.g. Warren et al. 2012)), and are also the more flare-productive, so this is an example of "large active region bias".

However, the $\mu(I_{\square})$ parameters perform poorly across wavelengths and event definitions: see for example $\mu(I_{94})$ for C1.0+/24 hr in Figure 4. The distributions are distinguishable (the means are separated), and the "event" distribution tends toward higher values, but there is no discriminant boundary within the bulk of the data. Pairing of these results (the $\Sigma(I_{\square})$ and $\mu(I_{\square})$ parameter performances), and looking in detail at the distributions, confirms that while in fact the $\mu(I_{\square})$ values are higher for the event populations, it is by not enough so as to provide good predictive power due to the class imbalance.

In other words, flare-imminent regions are inherently only slightly brighter (higher specific intensity) than flare-quiet regions. This result is a bit surprising, as one might expect that the magnetic complexity strongly related to flare productivity would produce strong corona-threading currents available to heat and preferentially brighten flare-imminent regions significantly over similarly-sized but flare-quiet sunspot groups (see, *e.g.* Asgari-Targhi et al. 2019). Such does not appear to be the case.

However, small structures that produce intense brightness variations are more likely to impact distribution wings. This can explain the dominating performances of parameters based on the kurtosis of the running-difference distributions ($\kappa(\Delta I_{\square})$; Tables 3-6 and Figures 7 – 10, see also Sections 4.2, 5.1). The ability of the $\kappa(\Delta I_{\square})$ parameters to distinguish flare-imminent from flare-quiet targets

indicates that flare-imminent regions display rapid variability in the E/UV images that is small in spatial scale, as well. These results are consistent with an increased number of small-scale ongoing reconnection events related to the increased magnetic complexity of these regions relative to AARPs that are imminently flare-quiet.

The skew of the running-difference has discriminating power for many filters and more than one event definition. Noting that, for example, 171A filter images are often used to detect coronal loops, we examined the parameter distributions and in fact the data display both positive and negative skew, but the positive-skew dominates and is slightly more pronounced for the event populations. In this context, the lack of performance for $\mu(\Delta I_?)$ implies that overall, the brightness changes on short timescales sum to zero. Hence, $\varsigma(\Delta I_?) > 0$ implies a small number of intense brightenings probably combined with a larger number of less intense dimmings to produce an imbalance in the wings of the running-difference brightness distributions.

Because we do not (yet) analyze the AARP data specifically in the context of, for example, nearby open magnetic flux, we cannot comment on whether we are detecting "crinkles" specifically (Sterling & Moore 2001b) or more generic enhanced small-scale activity. However, we can conclude that these results, based on moment analysis of time-series data, is likely only available because we are using full-resolution spatial sampling.

### 5.3. *Wavelength, Temperature, and Physical Regimes*

The AIA filters do not uniquely sample single temperatures or physical regimes (Lemen et al. 2012; Warren et al. 2012; Cheung et al. 2015). This fact makes direct interpretation of the parameters in the context of plasma temperature quite challenging if not potentially misleading, and obviating the need for, *e.g.*, differential emission measure analysis (forthcoming, see Section 6). Still, analysis of the results as a function of filter (Figures 7–10) shows patterns of behavior that are notable in the context of the different regimes that the filters do sample.

For some filters there is a significant difference in the BSS across event definitions between the $\Sigma(I_?)$ parameters and the higher-order $M(I_?)$ parameters. This result implies that the presence of emission is discriminating, but there is no further information from the spatial distribution of the emission. This trend is notably present in the 131, 171, 211, and 193 and 335 A filter results. In contrast, for the 94, 304, and 1600 A filters there is non-negligible performance for the higher-order $M(I_?)$ parameters in addition to the $\Sigma(I_?)$, implying that distinctive information about the spatial distribution (features) can be present. The common theme between the first set of filters is that they are sensitive to hotter plasma than are the 304, and 1600 A filters (Lemen et al. 2012; O'Dwyer et al. 2010). These two filters are not sensitive to flare-temperature plasma and while the 94A filter intensity is in fact dominated by hot plasma, it does include a cooler component (Warren et al. 2012).

The presence of hot plasma overall may be indicative of *past* flare activity, and we must be reminded that data acquisition is not separate from flare events (Figure 3). The presence of a single flare does not usually directly impact (for example) the inferred longer temporal behavior as parametrized by the "d/dt" variables (see Figure 2), although as mentioned earlier it can supply outlier events. But the $d\Sigma(I_?)/dt$ parameters stand out as well as the $\Sigma(I_?)$ parameters. To the extent that the images in the filters that may be dominated by flare-temperature plasma, this result signifies that its presence is an indicator of past and, hence, *future* activity. This result is reminiscent of the strong performance (reflected in "observer's wisdom") of "persistence" as a flare predictor (see discussions in Sawyer et al. 1986; Leka et al. 2019b)).

In line with this finding, we also see that parameters from the 94 and 131 Å filters, sensitive to hot flare plasma, dominate the top-performing parameters for the M1.0+ event definitions compared to C1.0+. This result implies that there is increased activity/dynamics in hotter plasma prior to M1.0+ flares, or that larger flares may be produced preferentially after smaller flares have energized the corona. However, we note that as the exact order of the top parameters for M1.0+/24 hr and especially M1.0+/6 hr is not very robust given the uncertainties, this result only provides a hint at the importance of the hottest channels for differentiating larger-flare-imminent regions.

Additionally, in some AIA filters and across event definitions (see Figures 7 – 10), the mean intensity $\mu(I_{\textcircled{?}})$ does not predict between the two populations well, but the standard deviation $\sigma(I_{\textcircled{?}})$ does. The spatial variation of the brightness is broader (larger standard deviation) for flare-imminent regions. For a few filters, notably 94, 304, and 1600A, this disparity extends to the higher-order moments of the intensity distribution, with notably better performance by $\varsigma(I_{\textcircled{?}})$ and $\kappa(I_{\textcircled{?}})$ than $\mu(I_{\textcircled{?}})$.

Two of those latter filters are distinctly *not* sensitive to flare-temperature coronal plasma. He II 304 Å is a relatively cool optically thick line sensitive to the chromosphere / upper transition region, with a peak temperature response around 0.05MK, albeit with challenging radiative transfer characteristics (Golding et al. 2017). It samples a different physical regime than the other filters which image the upper corona (see Figure 1), especially in the context of flares. The C IV and "continuum" 1600 Å filter samples the upper photosphere and transition region. While flare ribbons are often traced using 1600 Å emission, that emission is not particularly hot (Simões et al. 2019) – but the brightness in 1600A filter images is also sensitive to the presence of magnetic structures and localized areas of transient heating. The 94 Å filter images are generally dominated by hot active-region core plasma and flare plasma (Lemen et al. 2012; Cheung et al. 2015), but include a cooler-plasma component (Warren et al. 2012), and additionally have a notoriously low signal-to-noise ratio.

From all of this we can conclude that there is evidence of a characteristic difference in the distribution of intensity between flare-imminent and flare-quiet active regions. In the high corona, the features are more likely larger-scale, detectable by the standard deviation of the distribution, whereas in the upper photosphere, transition region, and chromosphere, the features are likely to include smaller-scale features that impact the higher-order moments.

The temporal evolution of the moments of the brightness distributions, also shows notable differences in patterns between filters that follow the same trends as outlined above: $d\mu(I_{\textcircled{?}})/dt$ shows no predictive capability across wavelength and event definition, $d\sigma(I_{\textcircled{?}})/dt$ only for 94, 304 and to a small extent 335A, then $d\varsigma(I_{\textcircled{?}})/dt$, $d\kappa(I_{\textcircled{?}})/dt$ show predictive power for 94, 304, and 1600A but not for the other filters. Again, this implies we detect evolution in the level of variability of small-scale intensity changes, as could be related to general magnetic complexity and associated on-going small reconnection events in the transition region and chromosphere. This variability is not reflected in parameters derived from filters that sample only hotter plasma, meaning we detect variations that are dominated by larger, less impulsively-varying structures.

The overall less-good performance of the 1600Å parameters across event definitions, specifically the $d\varsigma(\Delta I_{1600})/dt$ and $d\kappa(\Delta I_{1600})/dt$ compared to the strong results for the same parameters from filters that sample coronal heights and temperatures, strengthens the case that parameters using EUV filters detect small-scale reconnection events. Such phenomena may be insufficiently large or energetic enough to produce UV-radiation signatures in the lower layers of the solar atmosphere. At the chromospheric height and temperatures detected in the 304A channel, however, and the

higher / hotter channels, these small-scale high-frequency variations are visible and bring power to differentiating between the populations, across event definitions.

There are patterns in the results (Figures 7–10) which imply that the filters sensitive to more than one temperature detect different behaviors from the different physical regimes they sample. For example, filters that are predominantly sensitive to active-region plasma temperatures (171, 193, 211 Å) show poor performance for $\sigma(I_{\square})$ whereas 1600 Å shows moderate performance in that parameter, as do 94 Å and 304 Å. Similar behavior is seen for $d\kappa(I_{1600})/dt$, whereas $d\kappa(I_{171,193,211})/dt$ show poor performance. In contrast, the active-region plasma-dominated filters show moderate performance in $\sigma(\Delta I_{171,193,211})$ whereas $\sigma(\Delta I_{1600})$ does not. The 131 Å filter senses emission from both flare-relevant Fe XVIII but also cooler transition-region Fe VIII; the 304 Å line samples a mix of regimes; the 94 Å filter is sensitive to the transition-region sensing Fe IX, Fe X emission as well as the flare-relevant Fe XVIII. The performance patterns for the 94 Å filter parameters, as compared to those from the more selective hot- *vs.* cool-sensing filters, confirms that both flare- and transition-region behaviors are being detected in the 94 Å filter, especially as we have not corrected for the "warm" component (c.f. Warren et al. 2012). The dominance of the 94 Å filter parameters in overall performance shows that multi-regime sampling may enhance the breadth of information available on the flare-imminent nature of solar active regions.

This analysis of NPDA results for the AIA filters and the implied physical regimes they sample is not straightforward, that is very clear. Rather than pushing the analysis further with regards to physical interpretation, we acknowledge the need for, e.g., Differential Emission Measure analysis, which is beyond the scope of this article.

## 6. DISCUSSION

We present here a large-sample statistical analysis of the behavior of the solar chromosphere and corona as deduced from the parametrization of UV and EUV images from AIA. We specifically ask how these parametrizations behave in flare-imminent active regions. This study complements previous work that focuses on the photospheric magnetic field (Leka & Barnes 2007; Leka et al. 2018); we find that there is some information available to statistically, but not uniquely, differentiate between regions that will produce a flare event, according to various event definitions, from those that will not.

Superficially, the work by Nishizuka et al. (2017); Jonas et al. (2018); Alipour et al. (2019) appears similar to the present study, given their use of AIA data in the context of flare prediction. However, there are very important differences. First and foremost, this is *not* a study focused on empirical flare prediction, but rather we ask whether there are physical characteristics of flare-imminent active regions as viewed from chromospheric, transition region, and coronal emission. The data handling and preparation is different, performed here with a strong emphasis on ensuring the ability to perform quantitative physical analysis (Dissauer et al. 2022c). Lastly and most importantly, by constructing the parameters specifically to investigate physical behavior, including behavior on different temporal scales, the results can lead to some physical interpretation.

The results show classification performance that varies from "very good" through "mediocre" to "poor", depending on which combination of event definition and metric is used. The BSS is similar to what is achieved on similar-sized datasets when the question is posed for parametrizations of the photosphere; this metric provides a summary of how well the predicted probability for any given target reflects the frequency of occurrence for other samples with the same measure. High BSS is

extremely difficult to achieve as it is constructed against the climatology, and class-imbalance – while inconvenient, is a strong influence for this metric. It is a metric based on the probabilities and thus the true distribution of the parameters, so that the "mediocre" and worse scores reflect the fact that substantial differences in the distributions can be partially offset by the prior probabilities (Figures 4, 11).

The Max(TSS) results are good, but caution must be used to understand that this metric is optimized when the probability threshold used (or incorporated into a cost function, for example) reflects the event rate, again coming up against the class-imbalance reality of the Sun (Bloomfield et al. 2012; Barnes et al. 2016; Kubo 2019). Comparing the present results to the very similar targets (although different latency periods), sample sizes, and approach in Leka et al. (2018), the Max(TSS) results are similar even though that study invoked multi-parameter NPDA.

The impressive scores here are the ROCSS or G, which summarize the ROC plots and the correspondence between the value of a parameter, its associated probability, and whether or not there was a corresponding event. In this sense, we can definitively say that there is information in the coronal images that is related to whether or not a region produces a flare event as we define one, given the parameters we use.

As the event rate decreases (Table 2), the best BSS values get smaller while the Max(TSS), ROCSS, and G values get larger. The distributions of event-imminent versus event-quiet populations become increasingly different with lower event rates, which is reflected in the Max(TSS), ROCSS, and G values, but this is more than offset by the increasing class imbalance that enters into the BSS. Similar behavior is also present in predictions made from parameters characterizing the photosphere (Barnes et al. 2016; Leka et al. 2018, 2019a). Clearly, no single metric provides a thorough evaluation of performance, and factors such as class imbalance or event rate must also be considered when interpreting metrics, especially those for which thresholds or limits must be set.

We find that enhanced variability in EUV and UV intensity on short timescales and small spatial scales is one of the strongest discriminators across event definitions and AIA filters. This enhancement is most likely of the form of intense transient brightenings, whether small-scale and localized or rapid larger loop movement, rather than gradual loop movement or gradual heating/cooling, as it preferentially enhances the wings (extremes) of the running-difference image brightness distributions. Of note here, spatial resolution matters in order for the parametrizations to detect these differences, and these results validate our approach of retaining the full AIA spatial sampling across the AARP fields of view (Dissauer et al. 2022c).

On longer timescales, strong increases (or decreases) in brightness moderately indicates impending flaring, and while overall the presence of hot plasma is a good indicator, this result is also consistent with the general correlation between active region size and flare productivity. The evolution of parameters describing the corona can provide flare-imminent indicators, but with little preference over "static" parameters.

Of note, while coronal loop structures are readily detected through an analysis of the spatial variations of emission in the 171, 211 Å filters. the quantitative measure of these spatial variations (*e.g.* $\sigma(I_{171,211})$) is not a good discriminator. Also surprisingly poor is the mean intensity and its longer-term trending, which implies that there is minimal significant difference between magnetically complex and magnetically simple active regions in terms of their average coronal brightness and its temporal variation.

The differences in coronal, transition-region, and chromospheric E/UV emission between flare-imminent and not-flare-imminent active regions has broad implications for models of active-regions overall, and their upper atmospheres in particular. The approach outlined here and these results provide constraints on the expected emission and kinematic behavior of pre-event (and even post-event) active region upper atmospheres.

As pointed out in Section 5.3, simply analyzing the behavior of the brightness and kinematics in AIA filters is tricky due to their multi-thermal sensitivity. We address this in an upcoming work that uses differential emission measure analysis to disentangle densities and temperatures across this AARP database (Dissauer et al. 2022a). Similarly, a more complete picture will be built as we combine the AARP database with the HARP magnetic field inputs; as of this work we simply begin the process of statistically understanding the behavior of the chromospheric, transition region, and coronal regimes in the context of flare events using large-sample data finally afforded by high-resolution continual imagery from *SDO*/AIA.

## ACKNOWLEDGMENTS

The authors thank the referee for a thorough reading and insightful feedback that helped improve the paper. This work was made possible by funding primarily from NASA/GI Grant 80NSSC19K0285 with some initial exploration through AFRL SBIR Phase-I contract FA9453-14-M-0170, and some final support from NASA/GI Grant 80NSSC21K0738 and NSF/AGS-ST Grant 2154653.

*Facility:* SDO (HMI and AIA), GOES (XRS)

*Software:* SolarSoft (Freeland & Handy 1998), NCI (Leka et al. 2018)

## REFERENCES

Alipour, N., Mohammadi, F., & Safari, H. 2019, ApJS, 243, 20, doi: 10.3847/1538-4365/ab289b

Argiroffi, C., Peres, G., Orlando, S., & Reale, F. 2008, A&A, 488, 1069, doi: 10.1051/0004-6361:200809355

Asgari-Targhi, M., van Ballegooijen, A. A., & Davey, A. R. 2019, ApJ, 881, 107, doi: 10.3847/1538-4357/ab2e01

Bamba, Y., Kusano, K., Imada, S., & Iida, Y. 2014, PASJ, 66, S16, doi: 10.1093/pasj/psu091

Barnes, G., Birch, A. C., Leka, K. D., & Braun, D. C. 2014, ApJ, 786, 19, doi: 10.1088/0004-637X/786/1/19

Barnes, G., & Leka, K. D. 2006, ApJ, 646, 1303, doi: 10.1086/504960

Barnes, G., Leka, K. D., Schumer, E. A., & Della-Rose, D. J. 2007, Space Weather, 5, 9002, doi: 10.1029/2007SW000317

Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, ApJ, 829, 89, doi: 10.3847/0004-637X/829/2/89

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJL, 747, L41, doi: 10.1088/2041-8205

Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135, doi: 10.1088/0004-637X/798/2/135

Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549, doi: 10.1007/s11207-014-0529-3

Breiman, L. 2001, Machine Learning, 45, 5, doi: 10.1023/A:1010933404324

Cheung, M. C. M., Boerner, P., Schrijver, C. J., et al. 2015, ApJ, 807, 143, doi: 10.1088/0004-637X/807/2/143

Cho, K., Lee, J., Chae, J., et al. 2016, SoPh, 291, 2391, doi: 10.1007/s11207-016-0963-5

Cinto, T., Gradvohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. 2020, SoPh, 295, 93, doi: 10.1007/s11207-020-01661-9

Dissauer, K., Leka, K. D., Barnes, G., & Wagner, E. L. 2022a, ApJ, in preparation

Dissauer, K., Leka, K. D., & Wagner, E. L. 2022b, The NWRA AIA Active Region Patch Database, V1, doi: TBD

Dissauer, K., Leka, K. D., Wagner, E. L., & Barnes, G. 2022c, ApJ, accepted

Efron, B., & Gong, G. 1983, Am. Stat., 37, 36

Freeland, S. L., & Handy, B. N. 1998, SoPh, 182, 497, doi: 10.1023/A:1005038224881

Garcia, H. A. 1994, SoPh, 154, 275, doi: 10.1007/BF00681100

Georgoulis, M. K., Bloomfield, D. S., Piana, M., et al. 2021, Journal of Space Weather and Space Climate, 11, 39, doi: 10.1051/swsc/2021023

Golding, T. P., Leenaarts, J., & Carlsson, M. 2017, A&A, 597, A102, doi: 10.1051/0004-6361/201629462

Harra, L. K., Matthews, S., Culhane, J. L., et al. 2013, ApJ, 774, 122, doi: 10.1088/0004-637X/774/2/122

Hills, M. 1966, J. R. Statist. Soc. B, 28, 1

Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483, doi: 10.1007/s11207-014-0516-8

Imada, S., Bamba, Y., & Kusano, K. 2014, PASJ, 66, S17, doi: 10.1093/pasj/psu092

Jolliffe, I. T., & Stephenson, D. 2012, Forecast Verification: A Practioner's Guide in Atmospheric Science, 2nd Edition (The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: Wiley), doi: 10.1002/9781119960003

Jonas, E., Bobra, M., Shankar, V., Todd Hoeksema, J., & Recht, B. 2018, SoPh, 293, #48, doi: 10.1007/s11207-018-1258-9

Joshi, B., Veronig, A. M., Lee, J., et al. 2011, ApJ, 743, 195, doi: 10.1088/0004-637X/743/2/195

Krista, L. D., & Chih, M. 2021, ApJ, 922, 218, doi: 10.3847/1538-4357/ac2840

Kubo, Y. 2019, Journal of Space Weather and Space Climate, 9, A17, doi: 10.1051/swsc/2019016

Leka, K. D., & Barnes, G. 2003a, ApJ, 595, 1277

—. 2003b, ApJ, 595, 1296

—. 2007, ApJ, 656, 1173, doi: 10.1086/510282

Leka, K. D., Barnes, G., & Wagner, E. L. 2018, Journal of Space Weather and Space Climate, 8, A25, doi: 10.1051/swsc/2018004

Leka, K. D., Dissauer, K., Barnes, G., & Wagner, E. L. 2022, Replication Data for Properties of Flare-Imminent versus Flare-Quiet Active Regions from the Chromosphere through the Corona II: NonParametric Discriminant Analysis Results from NCI, DRAFT VERSION, Harvard Dataverse, doi: 10.7910/DVN/WPN39J

Leka, K. D., Park, S. H., Kusano, K., et al. 2019a, ApJS, 243, 36, doi: 10.3847/1538-4365/ab2e12

—. 2019b, ApJ, 881, 101, doi: 10.3847/1538-4357/ab2e11

Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, SoPh, 275, 17, doi: 10.1007/s11207-011-9776-8

Li, J., Mickey, D. L., & LaBonte, B. J. 2005, ApJ, 620, 1092, doi: 10.1086/427205

Machol, J. 2022

Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634, doi: 10.1088/0004-637X/723/1/634

McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. 2018, Journal of Space Weather and Space Climate, 8, A34, doi: 10.1051/swsc/2018022

Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, ApJ, 835, 156, doi: 10.3847/1538-4357/835/2/156

O'Dwyer, B., Del Zanna, G., Mason, H. E., Weber, M. A., & Tripathi, D. 2010, A&A, 521, A21, doi: 10.1051/0004-6361/201014872

Panos, B., & Kleint, L. 2020, ApJ, 891, 17, doi: 10.3847/1538-4357/ab700b

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3, doi: 10.1007/s11207-011-9841-3

Plowman, J. 2016, Journal of Space Weather and Space Climate, 6, A8, doi: 10.1051/swsc/2016002

Qiu, J., & Cheng, J. 2017, ApJL, 838, L6, doi: 10.3847/2041-8213/aa6798

Raboonik, A., Safari, H., Alipour, N., & Wheatland, M. S. 2017, ApJ, 834, 11, doi: 10.3847/1538-4357/834/1/11

Sawyer, C., Warwick, J. W., & Dennett, J. T. 1986, Solar Flare Prediction (Boulder, CO: Colorado Assoc. Univ. Press)

Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, SoPh, 275, 207, doi: 10.1007/s11207-011-9834-2

Schonfeld, S. J., White, S. M., Hock-Mysliwiec, R. A., & McAteer, R. T. J. 2017, ApJ, 844, 163, doi: 10.3847/1538-4357/aa7b35

Seki, D., Otsuji, K., Isobe, H., et al. 2017, ApJL, 843, L24, doi: 10.3847/2041-8213/aa7559

Silverman, B. W. 1986, Density Estimation for Statistics and Data Analysis (London: Chapman and Hall)

Simões, P. J. A., Reid, H. A. S., Milligan, R. O., & Fletcher, L. 2019, ApJ, 870, 114, doi: 10.3847/1538-4357/aaf28d

Sterling, A. C., & Moore, R. L. 2001a, ApJ, 560, 1045, doi: 10.1086/322241

—. 2001b, J. Geophys. Res., 106, 25227

Sterling, A. C., Moore, R. L., & Freeland, S. L. 2011, ApJL, 731, L3, doi: 10.1088/2041-8205/731/1/L3

Viall, N. M., & Klimchuk, J. A. 2012, ApJ, 753, 35, doi: 10.1088/0004-637X/753/1/35

Viereck, R. A., & Machol, J. L. 2017, in AGU Fall Meeting Abstracts, Vol. 2017, SH42A–06

Warren, H. P., Winebarger, A. R., & Brooks, D. H. 2012, ApJ, 759, 141, doi: 10.1088/0004-637X/759/2/141

Welsch, B. T., Li, Y., Schuck, P. W., & Fisher, G. H. 2009, ApJ, 705, 821, doi: 10.1088/0004-637X/705/1/821

Woods, M. M., Harra, L. K., Matthews, S. A., et al. 2017, SoPh, 292, 38, doi: 10.1007/s11207-017-1064-9

Zhang, Q. M., Su, Y. N., & Ji, H. S. 2017, A&A, 598, A3, doi: 10.1051/0004-6361/201629477