The Astrophysical Journal, 656:1173–1186, 2007 February 20

PHOTOSPHERIC MAGNETIC FIELD PROPERTIES OF FLARING VERSUS FLARE-QUIET ACTIVE REGIONS. IV. A STATISTICALLY SIGNIFICANT SAMPLE

K. D. Leka and G. Barnes
Colorado Research Associates Division, NorthWest Research Associates, Inc., Boulder, CO;
leka@cora.nwra.com, graham@cora.nwra.com
Received 2006 July 21; accepted 2006 October 16

ABSTRACT

Statistical tests based on linear discriminant analysis are applied to numerous photospheric magnetic parameters, continuing toward the goal of identifying properties important for the production of solar flares. For this study, the vector field data are University of Hawai'i Imaging Vector Magnetograph daily magnetograms obtained between 2001 and 2004. Over 1200 separate magnetograms of 496 numbered active regions comprise the data set. At the soft X-ray C1.0 level, 359 magnetograms are considered "flare productive" in the 24 hr postobservation. Considering multiple photospheric variables simultaneously indicates that combinations of only a few familiar variables encompass the majority of the predictive power available. However, the choice of which few variables is not unique, due to strong correlations among photospheric quantities such as total magnetic flux and total vertical current, two of the most powerful predictors. The best discriminant functions result from combining one of these with additional uncorrelated variables, such as measures of the magnetic shear, and successfully classify over 80% of the regions. By comparison, a success rate of approximately 70% is achieved by simply classifying all regions as "flare quiet." Redefining "flare-productive" at the M1.0 level, parameterizations of excess photospheric magnetic energy outperform other variables. However, the uniform flare-quiet classification rate is approximately 90%, while incorporating photospheric magnetic field information results in at most a 93% success rate. Using nonparametric discriminant analysis, we demonstrate that the results are quite robust. Thus, we conclude that the state of the photospheric magnetic field at any given time has limited bearing on whether that region will be flare productive.

Subject headings: methods: statistical — Sun: activity — Sun: flares — Sun: magnetic fields — Sun: photosphere

1. INTRODUCTION

What constitutes the difference between those solar active regions that produce energetic events and those that do not? The answer no doubt lies in the state and ongoing evolution of the magnetic field both in the immediate area of the active region and in the context of the nearby and overlying magnetic structure.

The search for distinguishing characteristics of a flare-productive active region has a long history. With several solar activity cycles providing consistent, modern data, a number of studies have investigated the statistical relation between active-region magnetic morphologies and flare productivity. Numerous different analysis methods have been considered, relying upon continuum light images, line-of-sight magnetic field maps, coronal morphology, or even helioseismology data. Highlights include investigations of flare productivity with magnetic "class" designation (McIntosh 1990; Bornmann & Shaw 1994), which is still the primary basis of most flare forecasting methods even when other information is considered, such as coronal morphology and line-of-sight magnetic morphology (NOAA/Space Environment Center [NOAA/SEC], and see Gallagher et al. 2002). Indeed, flare persistence can statistically provide a flare forecast of similar quality to that currently provided by NOAA/SEC, with no additional information concerning the details of visible active regions (Wheatland 2004, 2005). Still, new analysis of an old problem can be advantageous. Very recently, the flare productivity of active regions was compared to the power spectrum of magnetic features (Abramenko 2005), which indicated that the power-law index of a young active region may predict its later flare productivity. A fractal analysis of active regions (McAteer et al. 2005), also using line-of-sight magnetic field images, showed that there exist thresholds of fractal index below which active regions simply do not produce flares of a certain size. From helioseismology, a link has been shown between the subsurface flow patterns beneath active regions and flare productivity (Komm et al. 2005). In some recent studies, the "events" considered were coronal mass ejections rather than flares (Canfield et al. 1999; Falconer et al. 2003, 2006). While the two phenomena may differ in the details of energetic output and geo-effectiveness, we stress that the critical aspect is that these cited studies were carefully crafted to test the null hypothesis; that is, "control" data are included against which a correlation or trend can be tested.

In three papers of this present series, the question of identifying a flare-imminent solar active region has been approached in various ways. In Leka & Barnes (2003a, hereafter Paper I), time series of photospheric vector magnetic field data were introduced, and numerous parameters were derived from the spatial maps of the photospheric magnetic vector. A visual comparison of temporal variations in these parameters was performed for preflare and quiet times. While some parameters were observed to change uniquely during the preflare observations, the vast majority had no unique preflare signature easily discernible by eye. In Leka & Barnes (2003b, hereafter Paper II), a statistical approach based on discriminant analysis (DA) was introduced and applied to time series of photospheric vector magnetic field data for seven active regions comprising 10 "flaring" and 14 "flare-quiet" epochs. The results of this analysis included a demonstration that there exists no single parameter that reliably separated the samples of the two populations (i.e., "predicted" the upcoming flare without numerous false alarms), although examples were available of

perfect discrimination between the samples when multiple parameters were considered simultaneously. Still, the problem of small-number statistics loomed. Recently, in Barnes & Leka (2006, hereafter Paper III), we investigated whether information concerning the coronal topology, obtained from the application of a magnetic charge topology model (Barnes et al. 2005), could provide better discrimination when applied to the same samples (same photospheric time series) of the two populations. With the acknowledgment that small-number samples can still introduce statistical anomalies, it was shown that numerous parameters derived to describe the coronal topology and its temporal variation performed significantly better than parameters derived solely from the photospheric magnetic field.

In this, the fourth paper of the series, yet another approach is taken. The analysis focuses once again on parameters directly describing the maps of the photospheric magnetic vector; however, this time a different and much larger data set is tested. The focus shifts away from temporal variations that occur a short time prior to a solar flare and toward the analysis of daily samples of the two populations. By choosing to neglect temporal information, the available sample size is now increased to where some new statistical approaches are employed that were inappropriate for the smaller sample sizes considered in the previous papers.

2. DATA

The Imaging Vector Magnetograph (IVM) at the University of Hawai'i Mees Solar Observatory produces photospheric vector magnetic field maps with good spatial resolution and a field of view that routinely covers entire solar active regions. The general description of the instrument and data have been covered elsewhere (Mickey et al. 1996; LaBonte et al. 1999; LaBonte 2004; Leka & Barnes 2003a); thus, we focus here on the differing aspects of the present data set.

In its normal observing mode, the IVM rapidly repeats the data acquisition sequence on an active region selected by the observer with guidance from recent activity, NOAA/SEC activity forecasts, and external information such as the Max Millennium report; such time sequences characterized the data used in Papers I–III. Prior to this targeted "movie mode," however, the IVM obtains data in "survey" mode, acquiring the spectro-polarimetric data for single magnetograms of each active region on the visible solar disk. These survey data are subjected to a "quick-look" data reduction and inversion[1] and are saved locally. The quick-look data reduction is expected to show systematic effects due to the nature of the data reduction, such as saturation of the field strengths in sunspot umbrae and possibly a slight deflection of the azimuthal angle near umbral-penumbral boundaries, since magneto-optical effects are not accounted for. However, no such issue arising from the data reduction should preferentially affect one population over the other. As they are acquired in the morning, the seeing is generally good, and as temporal variations are not considered, the seeing of each individual magnetogram is not otherwise accounted for in the uncertainties. To demonstrate that these data were worthy of analysis, comparisons were made between quick-look and full reduction data for a handful of active regions, and this comparison is discussed in Appendix A.

The image-plane data were sampled with 1.1 arcsec^2 pixels and were automatically trimmed to a 192×192 pixel size to remove possible edge effects. The 180° ambiguity was resolved with two iterations of the University of Hawai'i approach, which minimizes the vertical current and the divergence of the magnetic field (Canfield et al. 1993; Metcalf et al. 2006): the first resolution is based on a linear force-free field constructed using the image-plane data and the force-free parameter α determined using the $\alpha_{\rm best}$ approach (Leka & Skumanich 1999), followed by a second iteration performed after redetermining α using the new heliographic results.

The survey data covering 2001–2004 initially included almost 4000 magnetograms. This time period was chosen with attention to providing a sufficient number of magnetograms during a time when the IVM instrument and quick-look data reduction did not change. For the analysis here, the data were further selected to only include single active regions within the trimmed field of view and to be free of visible defects (gross instrumental offsets or fringes in the magnetic field maps). Within a trimmed magnetogram, it was required that at least 64 data points exist for which both the line-of-sight field B_l and the transverse field B_t were greater than the 2 σ level; the general noise level for each component in each magnetogram is determined by integrating the histogram of field strengths in bright, quiet-Sun regions to 68% of the area under the curve. Magnetograms with the solar limb in the field of view were not included; generally, the minimum observing angle for the center of the field of view was $\mu = \cos\theta \approx 0.5$. As an additional precaution, however, a mask was implemented to zero out any pixels limbward of $\mu = \cos\theta < 0.25$; it was only invoked for a few magnetograms. All numbered active regions for which data were acquired when the IVM was observing were otherwise included: *no further selection for size, bipolar nature, complexity, or flaring history was imposed.* The final tally is 1212 magnetograms of 496 different active regions on 430 days.

The flare events were determined using the event logs for the *Geostationary Operational Environmental Satellite* (*GOES*) available through the National Geophysical Data Center.[2] All *GOES* soft X-ray events with a peak flux above 1.0×10^{-6} W m^{-2} (that is, a C flare or greater) that were also active region identified were tabulated for the 24 hr period after the acquisition of the region's magnetogram in the database. No region-associated Hα or radio-burst flares were included if they did not also register as a *GOES* event. There are no further distinctions here concerning the character of the flare.

For the results presented here, a region was classified as "flaring" in a Boolean manner if it produced at least one soft X-ray event of class T or greater in the 24 hr postmagnetogram and "flare-quiet" otherwise. For the majority of the results presented here, $T =$ C1.0, with a short discussion in § 4.3 where $T =$ M1.0. This fairly low threshold was chosen to balance statistical requirements with background X-ray contamination levels. There were indeed days when the background level was higher than C1.0, and as such there are undoubtedly flare-quiet regions that were misclassified as such, because they produced only a single small event that failed to register on the *GOES* list. However, there are numerous examples of *GOES* events in our database that have a peak soft X-ray flux below the calculated background level; the apparent contradiction is simply based on the details of how the daily background level is calculated. Thus, we argue that the number of misclassifications due to the *GOES* background level should be small. That is, while arguments can be made against the need to forecast small flares, we take the view that a small event is indicative of the same physics as prevails in larger events. With this definition, the database contains 359 magnetograms of flaring regions (29.6% of the total), with 111 of those having produced at least one M flare or greater (9.2%), and

[1] Images are available at http://www.solar.ifa.hawaii.edu/IVM/archive.html.

[2] See http://www.ngdc.noaa.gov.

TABLE 1

PARAMETERS USED IN THE DISCRIMINANT ANALYSIS

Description	Formula	Variable
Atmospheric Seeing		
Median of the granulation contrast	$s = \mathrm{median}(\Delta \mathrm{I})$	s
Distribution of Magnetic Fields		
Moments of vertical magnetic field	$B_z = \boldsymbol{B} \cdot \boldsymbol{e}_z$	$\mathcal{M}(B_z)$
Total unsigned flux	$\Phi_{\mathrm{tot}} = \sum \lvert B_z \rvert \, dA$	Φ_{tot}
Absolute value of the net flux	$\lvert \Phi_{\mathrm{net}} \rvert = \lvert \sum B_z \, dA \rvert$	$\lvert \Phi_{\mathrm{net}} \rvert$
Moments of horizontal magnetic field	$B_h = \left(B_x^2 + B_y^2 \right)^{1/2}$	$\mathcal{M}(B_h)$
Distribution of Inclination Angle		
Moments of inclination angle	$\gamma = \tan^{-1}(B_z / B_h)$	$\mathcal{M}(\gamma)$
Distribution of the Magnitude of the Horizontal Gradients of the Magnetic Fields		
Moments of total field gradients	$\lvert \nabla_h B \rvert = \left[(\partial B / \partial x)^2 + (\partial B / \partial y)^2 \right]^{1/2}$	$\mathcal{M}(\lvert \nabla_h B \rvert)$
Moments of vertical field gradients	$\lvert \nabla_h B_z \rvert = \left[(\partial B_z / \partial x)^2 + (\partial B_z / \partial y)^2 \right]^{1/2}$	$\mathcal{M}(\lvert \nabla_h B_z \rvert)$
Moments of horizontal field gradients	$\lvert \nabla_h B_h \rvert = \left[(\partial B_h / \partial x)^2 + (\partial B_h / \partial y)^2 \right]^{1/2}$	$\mathcal{M}(\lvert \nabla_h B_h \rvert)$
Distribution of Vertical Current Density		
Moments of vertical current density	$J_z = C(\partial B_y / \partial x - \partial B_x / \partial y)$	$\mathcal{M}(J_z)$
Total unsigned vertical current	$I_{\mathrm{tot}} = \sum \lvert J_z \rvert \, dA$	I_{tot}
Absolute value of the net vertical current	$\lvert I_{\mathrm{net}} \rvert = \lvert \sum J_z \, dA \rvert$	$\lvert I_{\mathrm{net}} \rvert$
Sum of absolute value of net currents in each polarity	$\lvert I_{\mathrm{net}}^B \rvert = \lvert \sum J_z (B_z > 0) \, dA \rvert + \lvert \sum J_z (B_z < 0) \, dA \rvert$	$\lvert I_{\mathrm{net}}^B \rvert$
Moments of vertical heterogeneity current density[a]	$J_z^h = C(b_y \partial B_x / \partial y - b_x \partial B_y / \partial x)$	$\mathcal{M}(J_z^h)$
Total unsigned vertical heterogeneity current	$I_{\mathrm{tot}}^h = \sum \lvert J_z^h \rvert \, dA$	I_{tot}^h
Absolute value of net vertical heterogeneity current	$\lvert I_{\mathrm{net}}^h \rvert = \lvert \sum J_z^h \, dA \rvert$	$\lvert I_{\mathrm{net}}^h \rvert$
Distribution of Twist Parameter		
Moments of twist parameter[b]	$\alpha = C J_z / B_z$	$\mathcal{M}(\alpha)$
Best-fit force-free twist parameter[b]	$\boldsymbol{B} = \alpha_{\mathrm{ff}} \nabla \times \boldsymbol{B}$	$\lvert \alpha_{\mathrm{ff}} \rvert$
Distribution of Current Helicity		
Moments of current helicity[c]	$h_c = C B_z (\partial B_y / \partial x - \partial B_x / \partial y)$	$\mathcal{M}(h_c)$
Total unsigned current helicity	$H_c^{\mathrm{tot}} = \sum \lvert h_c \rvert \, dA$	H_c^{tot}
Absolute value of net current helicity	$\lvert H_c^{\mathrm{net}} \rvert = \lvert \sum h_c \, dA \rvert$	$\lvert H_c^{\mathrm{net}} \rvert$
Distribution of Shear Angles		
Moments of 3D shear angle[d]	$\Psi = \cos^{-1}(\boldsymbol{B}^p \cdot \boldsymbol{B}^o / B^p B^o)$	$\mathcal{M}(\Psi)$
Area with shear $>\Psi_0$, $\Psi_0 = 45^\circ$, 80°	$A(\Psi > \Psi_0) = \sum_{\Psi > \Psi_0} dA$	$A(\Psi > 45^\circ)$, $A(\Psi > 80^\circ)$
Moments of neutral line shear angle	$\Psi_{\mathrm{NL}} = \cos^{-1}(\boldsymbol{B}_{\mathrm{NL}}^p \cdot \boldsymbol{B}_{\mathrm{NL}}^o / B_{\mathrm{NL}}^p B_{\mathrm{NL}}^o)$	$\mathcal{M}(\Psi_{\mathrm{NL}})$
Length of neutral line with shear $>\Psi_0$	$L(\Psi_{\mathrm{NL}} > \Psi_0) = \sum_{\Psi_{\mathrm{NL}} > \Psi_0} dL$	$L(\Psi_{\mathrm{NL}} > 45^\circ)$, $L(\Psi_{\mathrm{NL}} > 80^\circ)$
Moments of horizontal shear angle[e]	$\psi = \cos^{-1}(\boldsymbol{B}_h^p \cdot \boldsymbol{B}_h^o / B_h^p B_h^o)$	$\mathcal{M}(\psi)$
Area with horizontal shear $>\psi_0$	$A(\psi > \psi_0) = \sum_{\psi > \psi_0} dA$	$A(\psi > 45^\circ)$, $A(\psi > 80^\circ)$
Distribution of Photospheric Excess Magnetic Energy Density		
Moments of photospheric excess magnetic energy density[d]	$\rho_e = (\boldsymbol{B}^p - \boldsymbol{B}^o)^2 / 8\pi$	$\mathcal{M}(\rho_e)$
Total photospheric excess magnetic energy	$E_e = \sum \rho_e \, dA$	E_e

NOTES.—The $\mathcal{M}(x)$ denotes taking the first four moments of the distribution of the variable x: the mean $\bar{x}$, the standard deviation $\sigma(x)$, the skew $\varsigma(x)$, and the kurtosis $\kappa(x)$. The C indicates physical constants that are included in the calculation but not listed here for clarity.

[a] Zhang (2001).

[b] Leka & Skumanich (1999).

[c] Abramenko et al. (1996); Bao et al. (1999).

[d] Wang et al. (1996).

[e] Hagyard et al. (1984), although B_h is used here, rather than $B_\perp$.

20 having produced at least one X flare (1.7%); no flares of at least C class were recorded for the remaining 853 magnetograms (70.4% of the total).

From each photospheric vector magnetic field map, most of the same parameters discussed in Papers I and II were constructed. There are essentially eight categories of variables, each evaluated over the field of view and describing the distribution of:

1. the magnetic fields B_z and B_h,
2. the inclination angle $\gamma = \tan^{-1}(B_z/B_h)$,
3. the horizontal gradients of the magnetic fields $|\nabla_h B|$, $|\nabla_h B_z|$, and $|\nabla_h B_h|$,
4. the vertical current density $J_z \sim \partial B_y/\partial x - \partial B_x/\partial y$,
5. the twist parameter $\alpha \sim J_z/B_z$,
6. part of the current helicity density $h_c \sim J_z B_z$,
7. the shear angle $\Psi = \cos^{-1}(\boldsymbol{B}^p \cdot \boldsymbol{B}^o/B^p B^o)$,
8. the photospheric excess magnetic energy density $\rho_e = (\boldsymbol{B}^p - \boldsymbol{B}^o)^2/8\pi$.

These are supplemented by a measure of the atmospheric seeing, which is included as a control variable that should have no predictive power. See Table 1 for a summary of the variables, and Paper I for detailed descriptions. For the present analysis, all variables were computed on a uniform image-plane grid, primarily due to issues of handling the large arrays that result from transforming off-disk-center data to the heliographic grid. Spatial derivatives were computed using heliographic coordinates, however, and the linear and areal dimensions of each pixel were computed so that all variables in fact result in physically relevant heliographic quantities.

For the sake of objectivity and considering the size of the database, it is essential to characterize the distributions of these variables in a completely autonomous manner. Thus, following Papers I–III, the distribution of a variable x is parameterized by its first four moments: mean $\bar{x}$, standard deviation $\sigma(x)$, skew $\varsigma(x)$, and kurtosis $\kappa(x)$ (Leka & Barnes 2003a, 2003b; Barnes & Leka 2006). The mean and standard deviation are familiar to most readers, giving the typical value of the distribution and the spread about that typical value. The skew describes the asymmetry of the distribution, indicating the presence of a one-sided tail. The kurtosis is normalized to zero for a Gaussian distribution, and deviations from zero indicate whether the distribution has long or short tails in comparison to a Gaussian distribution. The skew and kurtosis are sensitive to small patches of extreme values. Thus, for example, a highly twisted δ-spot in an otherwise potential region should appear as a significant nonzero skew in the distribution of α. In some cases, the moments are supplemented by the total and/or net value of the variable; when considering the distribution of the magnetic shear angle, we also consider the total area of strong shear and the length of strongly sheared neutral lines, as originally proposed by Hagyard et al. (1990) and incorporated into more recent studies (e.g., Falconer et al. 2006).

Because the variables are computed on a uniform grid in image coordinates, each moment is weighted by the pixel area, except the neutral line shear angle, which is weighted by a typical linear dimension for each pixel. To understand why this weighting is used, consider the weighted mean value of the vertical field,

$$\overline{B}_z = \frac{\sum B_z(\Delta\,\mathrm{Area})}{\sum(\Delta\,\mathrm{Area})} = \frac{\Phi_{\mathrm{net}}}{\mathrm{Area}}, \qquad (1)$$

where ΔArea is the area of a pixel. By weighting the vertical field by the area of a pixel, the first moment (the mean) represents the net flux divided by the total area, rather than being affected by variations in pixel area at different observing angles. This approach is equivalent to interpolating the observations to a regular grid in heliographic coordinates before taking the (unweighted) mean, as was done in Papers I and II. Unlike in Papers I and II, we do not consider *changes* in the parameters leading up to a flare, as only a single magnetogram characterizes each 24 hr period. Thus, the parameters described here are equivalent to the temporal means used in Papers I and II.

3. DISCRIMINANT ANALYSIS

To determine which properties of an active region are important for the occurrence of a flare, we use the same statistical approach as in previous papers of this series: discriminant analysis (e.g., Kendall et al. 1983; Anderson 1984). The basic goal of discriminant analysis is to classify a new measurement as belonging to one of two mutually exclusive groups, in this case flaring or flare quiet. This approach has the advantage of being able to simultaneously consider multiple parameters, so that if several conditions must be met before a flare can occur, discriminant analysis will be able to determine this.

Parameter space is divided into two regions such that measurements from a new magnetogram that fall in one region are predicted to flare, while measurements that fall in the other are predicted to be flare quiet. The boundary between the two is defined by where the discriminant function vanishes, and is constructed so as to maximize the overall rate of correct predictions. Assuming that the variables' distributions for both populations are Gaussian with equal covariance matrices, the discriminant function is linear in all the variables, and the boundary is a hyperplane, which is simply a line in two dimensions. Unlike in Paper II, it is assumed here that the a priori probability of membership in each population is proportional to the sample size. This results in a discriminant boundary that does not necessarily pass through the midpoint between the means of the samples, as had been the case in Paper II.

Discriminant analysis as implemented here minimizes the overall rate of misclassification. Particularly when the a priori probabilities of membership in the two populations are significantly different from 0.5, as is the case here, this can lead to a situation that favors errors of one type (the off-diagonal element in a classification table) over the other: for example, many more flaring data points can be predicted to be flare-quiet than flare-quiet data are predicted to flare. In comparison, the success rate quoted in Falconer et al. (2003, 2006) for a ± 2 day window assumes that there are equal errors of both types; that approach does not necessarily lead to the lowest overall error rate.

The performance of the discriminant function is typically judged by estimating the error rate, that is, the fraction of measurements that are incorrectly classified by the discriminant function. The standard and most straightforward way to estimate the error rate for a discriminant function is by way of a classification table in which the discriminant function is used to make a prediction about each of the data points used to construct it, and the prediction is then compared to what actually occurred. Unfortunately, this approach is biased (Hills 1966), simply because it judges the performance of the discriminant function based on the same data set used to construct it.

An unbiased estimate of the error can be obtained from the $n-1$ error rate (Hills 1966; Leka & Barnes 2003b), in which one data point is excluded and the discriminant function is constructed from the remaining $n-1$ data points. The discriminant function is then used to classify the excluded point, and the procedure is repeated for all n points. While this approach provides an unbiased estimate, it involves evaluating n discriminant functions to obtain one error rate, and with our sample size of

TABLE 2
10 Best-Performing Single-Variable Linear Discriminant Analysis Results

Mahalanobis Distance		Classification Table Error Rate		$n-1$ Error Rate	
Variable	D_M	Variable	Rate	Variable	Rate
$\Phi_{\rm tot}$	1.3132	$\Phi_{\rm tot}$	0.2277	$\Phi_{\rm tot}$	0.2277
$I_{\rm tot}$	1.0974	$I_{\rm tot}$	0.2318	$I_{\rm tot}$	0.2335
$I^h_{\rm tot}$	1.0499	$H_c^{\rm tot}$	0.2351	$H_c^{\rm tot}$	0.2351
$H_c^{\rm tot}$	0.9713	$\sigma(\Psi_{\rm NL})$	0.2384	$\sigma(\Psi_{\rm NL})$	0.2384
$\sigma(\Psi_{\rm NL})$	0.8958	$I^h_{\rm tot}$	0.2393	$I^h_{\rm tot}$	0.2393
E_e	0.7684	$I^B_{\rm net}$	0.2516	$I^B_{\rm net}$	0.2516
$\sigma(\rho_e)$	0.7435	$\sigma(\rho_e)$	0.2516	$\sigma(\rho_e)$	0.2516
$I^B_{\rm net}$	0.7235	$\lvert H_c^{\rm net}\rvert$	0.2550	$\lvert H_c^{\rm net}\rvert$	0.2550
$\overline{\Psi}$	0.7089	E_e	0.2566	E_e	0.2566
$\sigma(\Psi)$	0.6968	$L(\Psi_{\rm NL} > 45^\circ)$	0.2607	$L(\Psi_{\rm NL} > 45^\circ)$	0.2607

$n = 1212$, this rapidly becomes prohibitively slow. In practice, with the large sample sizes presented here, the bias introduced by using the classification table error rate is relatively small, but the $n-1$ error rate is also computed for key examples to ensure there are no large differences.

For most of the results presented here, it is assumed that the populations have Gaussian distributions with equal covariance matrices. This assumption is clearly not valid for all variables: for example, the total magnetic flux is a positive definite quantity, and so cannot have a truly Gaussian distribution. However, the error rates that result from this assumption can be considered lower bounds on the information content of the variables, as knowledge of the true distributions would lead to a better discriminant. This possibility is discussed in Appendix B in the context of first relaxing the assumption of equal covariance matrices and second in the context of nonparametric statistics. The results are surprisingly robust to the assumption of Gaussian distributions with equal covariance matrices.

The error rate of the discriminant function is an indication of the degree to which the two populations (flaring vs. flare quiet) overlap. However, with a very wide disparity between the sample sizes of the two populations, it is easy to construct hypothetical examples in which the probability density of one population falls wholly below the other, even though there is physically and statistically a significant difference between the two populations. In this case, it may be more meaningful to consider the Mahalanobis distance, which is a standardized measure of the distance between the sample means (Kendall et al. 1983). Combined with the number of degrees of freedom, the Mahalanobis distance can be compared to Hotelling's T^2 distribution to determine the probability that the null hypothesis can be rejected (Kendall et al. 1983), the null hypothesis being that the flaring and flare-quiet samples come from the same population (see also the Appendix in Leka & Barnes 2003b). From the point of view of understanding and modeling solar flares, it would be extremely interesting to identify variables that clearly discriminate between flare-imminent and flare-quiet states, so the focus here is on error rate estimates. The use of the Mahalanobis distance becomes most appropriate when we consider larger flares (§ 4.3).

Before proceeding, however, it is shown in Table 2 that the single-variable rankings are relatively insensitive to which measure is used to judge their performance: the Mahalanobis distance, the classification table error rate, or the $n-1$ error rate. The top two single variables, $\Phi_{\rm tot}$ and $I_{\rm tot}$, are ranked the same by all three measures, and most of the same variables appear in the top 10, although in slightly different order. The $n-1$ rate is the same as the classification table rate for all but one of the variables, indicating that the bias described earlier is small. It is worth noting that for the one exception, the $n-1$ error rate is larger than the classification table error rate, as the latter does tend to underestimate the true error rate. Given the consistency of these results, and particularly the small differences between the two error rate calculations for these data, the classification table approach is used hereafter to determine the best combinations of more than one variable. Additionally, for ease of comparison with other studies, the success rate (defined as 1.0—the classification table error rate) will henceforth be used and quoted as a percentage.

4. THE BEST VARIABLES AND THE BEST VARIABLE COMBINATIONS

Of particular interest for understanding the process that triggers solar flares is knowing what, if any, distinct configuration the photospheric magnetic field must assume for a flare to occur. Discriminant analysis can shed some light on this by determining the variables with the greatest ability to distinguish between a flare-producing and flare-quiet field. Further, it can determine the *combinations* of variables that are best able to identify a flare-productive magnetic field. However, there are a number of ways of selecting those variables, some impractical from a computational point of view. In this section, the optimal number of variables is determined first, then the best variables are presented.

4.1. *How Many Variables Are Needed?*

Perhaps the simplest approach for determining the best variables is to construct a single discriminant function of all the variables under consideration. When the variables are in standardized form, the magnitude of each variable's coefficient in the discriminant function gives its relative predictive power, provided the variables are not correlated (e.g., Klecka 1980). If some of the variables are correlated then the predictive power of correlated variables will be shared. The magnitude of the coefficient may not be particularly large for a variable with quite a large predictive capability because of the masking by a variable with which it is strongly correlated. Thus, any variable with a large coefficient is expected to have a relatively large amount of predictive power, but a small coefficient does not necessarily mean that a variable is without predictive power. Thus, considering the magnitudes of the standardized coefficients will yield at least some of the variables with the greatest predictive power, but it is not the best approach for strongly correlated variables.

Table 3 lists the best-performing variables based on their coefficients in an all-variable discriminant function. Several variables are also included that appear to perform less well but are

TABLE 3
All-Variable Discriminant Function Coefficients

Rank	Variable	Coefficient
1	$\Phi_{\rm tot}$	1.385
2	$I_{\rm tot}$	1.232
3	$I^h_{\rm tot}$	−1.230
4	E_e	−1.016
5	$\overline{\lvert\nabla_h B_z\rvert}$	0.848
6	$\sigma(\lvert\nabla_h B_z\rvert)$	0.730
13	$\sigma(\Psi_{\rm NL})$	0.402
16	$L(\Psi_{\rm NL} > 45^\circ)$	0.315
19	$H_c^{\rm tot}$	0.257
28	$\sigma(\rho_e)$	0.174

TABLE 4
Best m-Variable Combinations

m	Variables	Success Rate
1	$\Phi_{\rm tot}$	77.23
2	$\Phi_{\rm tot}, \sigma(\Psi_{\rm NL})$	79.37
3	$\Phi_{\rm tot}, \sigma(\Psi_{\rm NL}), \overline{\Psi}$	80.28
4	$\Phi_{\rm tot}, \sigma(\Psi_{\rm NL}), \sigma(\alpha), \sigma(\gamma)$	80.45
5	$\Phi_{\rm tot}, \sigma(\Psi_{\rm NL}), \kappa(\Psi_{\rm NL}), A(\psi > 45^\circ), \varsigma(\rho_e)$	80.69
74	all	81.02

of interest elsewhere. Figure 1 shows the magnitudes of all the coefficients; note from both the table and the figure that the magnitude of the coefficients drops off quite rapidly for the first seven variables and continues to decrease significantly for about another four variables. This indicates that larger combinations of variables are unlikely to greatly increase the performance of the discriminant function. The all-variable discriminant function has a success rate of 81.02%.

Another way to deal with issues surrounding correlated variables is to consider the discriminant functions of all permutations of a given number of variables. This is the most complete approach, but it suffers from serious computational drawbacks for the sample size and number of variables considered here. To determine the best m-variable combination out of a total of n_v variables requires evaluating $n_v!/(n_v - m)!$ permutations; for this investigation $n_v = 74$, so determining the best five-variable combination requires evaluating more than 10^9 discriminant functions. Thus, this approach is employed only for combinations of five or fewer variables. The results for discriminant functions of different numbers of variables are shown in Table 4. Most of the predictive power available in these data are realized by considering small numbers of variables. The best three-variable discriminant function has a 80.28% success rate, only 0.74% lower than the all-variable discriminant function, which should have the best success rate. Thus, using combinations with more variables improves the classification rate by less than a percent.

This is confirmed yet another way, a "step-up" procedure in which a single m-variable combination, starting with the best-performing single variable, is combined with each of the remaining variables in turn, and the best-performing $m + 1$ variable combination is retained (Klecka 1980). This approach is not guaranteed to determine the best-performing m-variable combination, but it is quick, and it produces a combination that performs very nearly as well as the best. In addition, the variable added with each step is unlikely to be correlated with recently added variables, since adding correlated variables will generally not significantly improve the discrimination. The step-up approach can be used to see how rapidly the predictive power changes with increasing numbers of variables. Figure 1 shows the results of the step-up procedure, specifically based on both the Mahalanobis distance and the classification error rate. It is evident from the figure that little independent information is gained after about the first three variables.

4.2. *What Are the Photospheric Characteristics of a Flare-imminent Region?*

All of the approaches considered indicate that a function of a few variables contains the majority of the discriminating power available. However, combining the two best-performing single variables in Table 5, $[\Phi_{\rm tot}, I_{\rm tot}]$, leads to a discriminant function with a success rate of only 77.64%, compared to the best two-variable

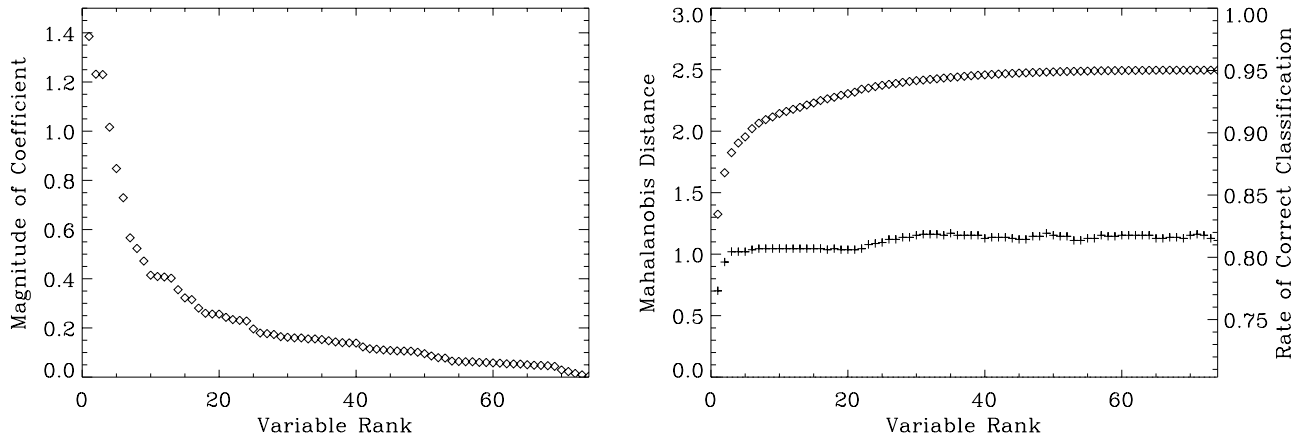

Fig. 1.—*Left*: Magnitudes of the coefficients in an all-variable discriminant function. For uncorrelated, standardized variables, the magnitudes of the coefficients indicate the relative predictive power of the variables. There is a rapid decrease in the magnitude of the coefficients for the first seven variables, and a further significant decrease for roughly the next four variables, but after about the first 10 variables, the coefficients decrease slowly, indicating that a discriminant function containing more variables will have little additional predictive capability. *Right*: Mahalanobis distance (*diamonds*), which indicates how likely the samples are from different populations, and the rate of correct classification (*plus signs*), based on the classification table. Both of these measures, resulting from the discriminant functions of m-variables (here m is the "variable rank"), confirm the results of the standardized coefficients shown at left: significant improvements occur for the first few variables, but little improvement occurs after about 10 variables.

TABLE 5
RANKING OF VARIABLES' PREDICTIVE POWER

One-Variable DF		Two-Variable DF		Four-Variable DF	
Variable	Success Rate	Variable	Frequency	Variable	Frequency
Φ_{tot}	77.23	Φ_{tot}	73	Φ_{tot}	62054
I_{tot}	76.82	I_{tot}	73	I_{tot}	52086
H_c^{tot}	76.49	I_{tot}^h	73	I_{tot}^h	39620
$\sigma(\Psi_{NL})$	76.16	H_c^{tot}	73	H_c^{tot}	37809
I_{tot}^h	76.07	$\sigma(\Psi_{NL})$	73	$\sigma(\Psi_{NL})$	29103
$\lvert I_{net}^B \rvert$	74.84	$\lvert I_{net}^B \rvert$	67	$\sigma(\Psi)$	11510
$\sigma(\rho_e)$	74.84	$\sigma(\rho_e)$	48	$\overline{\Psi_{NL}}$	11260
$\lvert H_c^{net} \rvert$	74.50	E_e	22	$\varsigma(B_h)$	10841
E_e	74.34	$\lvert H_c^{net} \rvert$	18	$\overline{\nabla_h B_z}$	10764
$L(\Psi_{NL} > 45^\circ)$	73.93	$\overline{\Psi_{NL}}$	13	$\overline{\psi_{NL}}$	10628

combination of $[\Phi_{tot}, \sigma(\Psi_{NL})]$, with a 79.37% success rate. The reason for this can be seen by considering the relationship between pairs of variables. Figure 2 shows the strong correlation between Φ_{tot} and I_{tot} compared to the lack of correlation between Φ_{tot} and $\sigma(\Psi_{NL})$. Combining Φ_{tot} and I_{tot} into a single discriminant function provides little additional information compared to a discriminant function of either variable alone, whereas including the combination of *either* Φ_{tot} *or* I_{tot} and $\sigma(\Psi_{NL})$ does improve the performance, because $\sigma(\Psi_{NL})$ is not strongly correlated with either Φ_{tot} or I_{tot}.

Because of the issue of correlated variables, it is possible that more than a few variables have significant predictive power. Indeed, there are a number of strongly correlated variables that have significant predictive power. To determine what parameters are best able to distinguish a flare-producing photospheric magnetic field, the approach described in Paper II is used: the discriminant functions are evaluated for all permutations of m-variables, and the best-performing variable combinations are retained (500 for $m = 2$; 200,000 for $m = 4$). The number of best-performing discriminant functions in which a particular variable appears is recorded and provides the basis for ranking that variable (Table 5). The number of top-performing combinations retained is chosen to sufficiently sample the available combinations. The one-variable ranking is based on the classification success rate. The exact order of the ranking is not maintained when varying m, but for $m \leq 5$ many of the same variables show up as being the best performing.

Of the top variables listed in Table 5, there are strong correlations among the total flux Φ_{tot}, the total vertical current I_{tot}, as well as the total heterogeneity current I_{tot}^h, the total current helicity H_c^{tot}, and the total excess energy E_e. Note that, in Table 3, the coefficient of Φ_{tot} is much larger than the coefficient of H_c^{tot} for the all-variable discriminant function. According to the rankings in Table 5, the two parameters actually have similar predictive power, a fact that is masked in the ranking by the all-variable coefficients due to the high correlations between the two. In addition, the standard deviation of the neutral line shear $\sigma(\Psi_{NL})$ is weakly correlated with a different group of variables. The best

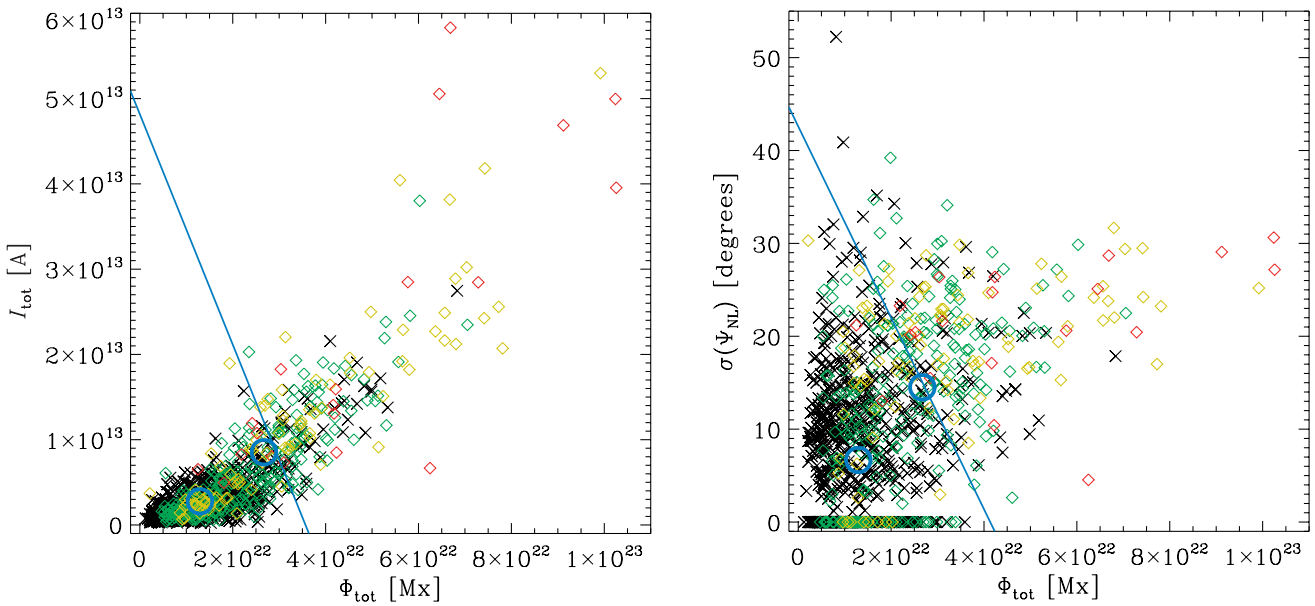

FIG. 2.—Examples of two-variable discriminant functions for the strongly correlated pair Φ_{tot} and I_{tot} (*left*) and the uncorrelated pair Φ_{tot} and $\sigma(\Psi_{NL})$ (*right*). Nonflaring regions (*crosses*) and flaring regions (*diamonds*) are shown with the largest flare in any 24 hr period (C, M, X) indicated by color (*green, yellow, and red, respectively*). The mean of each sample is shown (*blue circles*) as is the discriminant function (*blue line*). There are a number of points with $\sigma(\Psi_{NL}) = 0.0$ from regions where there are no well-measured horizontal fields on the neutral line. The points with $\Phi_{tot} \sim 10^{23}$ Mx are region NOAA AR 10486.

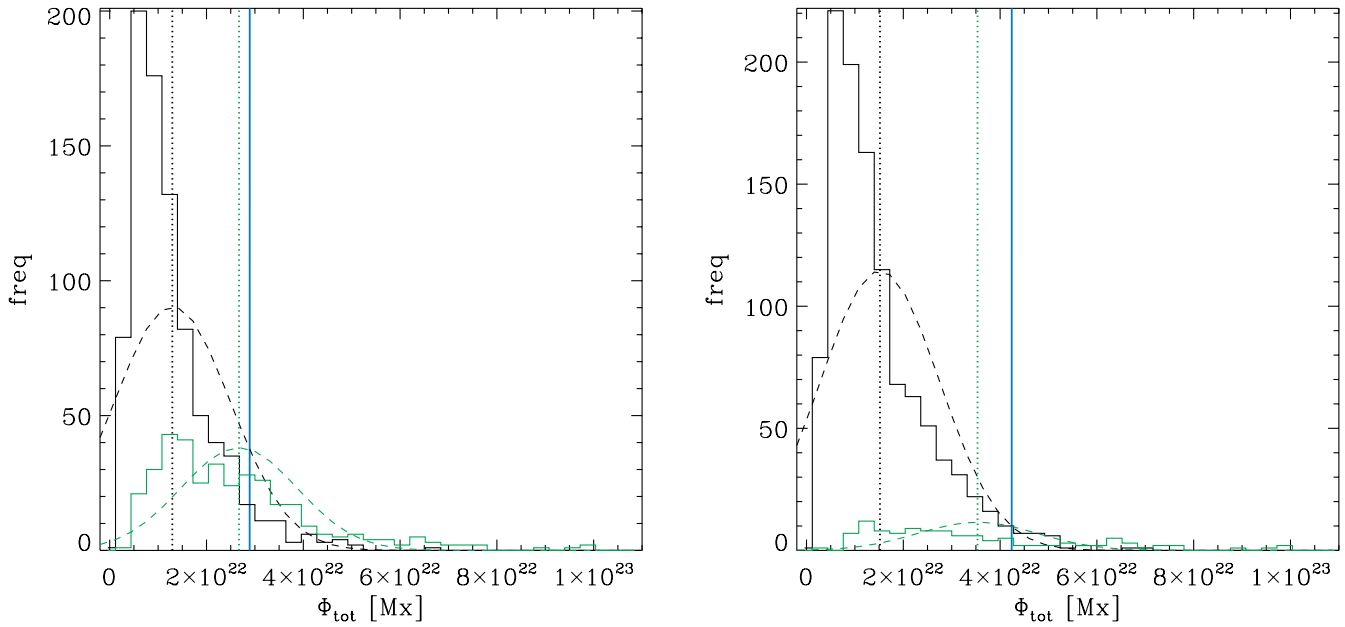

FIG. 3.— Comparison of linear discriminant functions for Φ_{tot} for a flare/flare-quiet threshold set at C1.0 (*left*) and M1.0 (*right*). The histograms show the flaring (*green*) and nonflaring (*black*) probability densities. The normal distribution fits to the data (*dashed lines*) and the means (*dotted lines*) are also shown. The point at which the two density estimates are equal is shown with a blue line that corresponds to the location of the discriminant function. The sample sizes differ significantly, as expected, yet the means are widely separated; the discriminant boundary lies out in the tails of the distributions, where the assumptions fail and the discriminant analysis performs poorly.

discriminant functions result from taking one variable from each of the correlated groups, rather than simply combining the best individual variables. Therefore, the regions that are most likely to be flare productive are those that are big (large total flux), which necessarily implies a large total current, large current helicity, etc., *and* those that have a large range of shear angles along their neutral lines, which typically implies other localized patches of strong shear.

These results are based on a linear discriminant. For most variables, a quadratic discriminant shows a very slight improvement over a linear discriminant, and a nonparametric discriminant shows only a slightly larger improvement. For example, the success rate for the total flux increases from 77.23% to 77.39%. For a few variables, most notably E_e, the improvement is much greater, from a linear discriminant success rate of 74.34% to a nonparametric success rate of 78.71%, the highest of any single variable, but still comparable to those variables with which it is correlated. Similarly, $L(\Psi_{NL} > 45^\circ)$ shows a significant improvement and is weakly correlated with $\sigma(\Psi_{NL})$ and thus may also be a good variable. For a detailed discussion of the nonparametric results, see Appendix B.

Using a nonparametric estimate for the probability density results in a large improvement in the ability of a few variables to discriminate between flaring and flare-quiet regions. However, there is only a slight improvement in the best single-variable nonparametric discriminant (as compared to the best single-variable linear discriminant), and there is still only a modest improvement in the highest success rate as compared with the prediction that nothing will flare. Estimating the probability density using nonparametric techniques requires *extremely* large sample sizes when considering multiple variables simultaneously. Thus, we cannot use the present sample to construct an all-variable nonparametric discriminant function. Based on the results for one- and two-variable discriminant functions, we believe that the improvements over linear discriminants will not be large.

The best variables listed in Table 5 have very little overlap with the best variables determined in Paper II. There are a number of reasons for this. Over half the best variables in Paper II describe the evolution of the photospheric magnetic field, which could not be calculated here. The time interval in question has been extended to 24 hr after the observation, compared to approximately 1 hr epochs in Paper II. Finally, the sample size in Paper II was too small to be statistically significant, so the results there were presented as a demonstration of the method. Thus, it is not surprising that a different set of variables performs well in this study.

4.3. *What Are the Photospheric Requirements for Large Flares?*

Heretofore, our approach has been to investigate what characteristics are required to produce rapid reconnection events (flares), regardless of the peak energy flux released (with the caveat of the soft X-ray background level). It is common to focus on what aspects of solar magnetic fields are required to specifically produce large flares.

With the database presented here, it is now possible to comment on the latter aspect using discriminant analysis by raising the minimum threshold of what is classified as flaring from C1.0 to M1.0. Only 9.2% of our full data set (111 magnetograms) are associated with the production of a flare event of size M1.0 or greater within the 24 hr after the magnetogram acquisition. In other words, a prediction that no region will ever produce a large flare should have a success rate of 90.84%.

This much smaller sample size of regions producing large flares implies that the a priori probability of belonging to the flaring population is much smaller. In this case, the probability of flaring is larger than that of being flare-quiet only far out in the tail of the distribution (see Fig. 3). Unfortunately, the tail of the distribution is where the Gaussian assumption is least appropriate. The tail of the distribution is also where the nonparametric estimates are least helpful, because there are too few data points to accurately reconstruct the true probability density function. The best discriminant functions make very small improvements to the success rate, with the best single variable, E_e, having a

TABLE 6
RANKING OF VARIABLES' MAHALANOBIS DISTANCE FOR M FLARES

ONE-VARIABLE DF		TWO-VARIABLE DF		FOUR-VARIABLE DF	
Variable	D_M	Variable	Frequency	Variable	Frequency
E_e	3.415	E_e	73	E_e	62196
$I_{\rm tot}$	3.363	$I_{\rm tot}$	73	$I_{\rm tot}$	56452
$I^h_{\rm tot}$	3.299	$I^h_{\rm tot}$	73	$I^h_{\rm tot}$	49993
$H_c^{\rm tot}$	3.101	$H_c^{\rm tot}$	73	$H_c^{\rm tot}$	34144
$\Phi_{\rm tot}$	2.700	$\Phi_{\rm tot}$	73	$\Phi_{\rm tot}$	18177
$\sigma(\rho_e)$	2.581	$\sigma(\rho_e)$	73	$L(\Psi_{\rm NL} > 45^\circ)$	12220
$L(\Psi_{\rm NL} > 45^\circ)$	2.466	$L(\Psi_{\rm NL} > 45^\circ)$	64	$\sigma(\rho_e)$	12210
$\lvert I^B_{\rm net}\rvert$	2.240	$\lvert I^B_{\rm net}\rvert$	19	$\overline{\rho_e}$	12085
$\overline{\rho_e}$	1.997	$\overline{\rho_e}$	12	$\sigma(\Psi)$	11241
$\lvert H_c^{\rm net}\rvert$	1.955	$\lvert H_c^{\rm net}\rvert$	11	$A(\psi > 45^\circ)$	10602

92.08% success rate, and even the best five-variable combination only achieving a 92.82% success rate.

With such small improvements in the success rates, can the two populations be distinguished? Yes, if the population means are well separated. Using the Mahalanobis distance rather than the success rate enables us to determine the probability that the null hypothesis (that the two samples come from the same population) can be rejected. For the sample sizes present here, a Mahalanobis distance greater than 0.4 gives a probability greater than 0.999999997; thus, hereafter solely the Mahalanobis distances are quoted for clarity.

The best variables for distinguishing regions capable of large flares, as determined by the probability of rejecting the null hypothesis, are shown in Table 6. It is interesting that the best-performing single variable is E_e, or the total "excess" magnetic energy density, with a Mahalanobis distance of 3.415. The best m-variable discriminant functions up to $m = 5$ all include E_e, and several other moments of the excess magnetic energy are present in the top 10 that were not common for the C1.0 threshold. This result implies in a fairly direct manner (rather than quite indirectly, as in Falconer et al. 2006) that active regions with substantial magnetic free energy are more likely to produce large solar flares.

5. CONCLUSIONS

By applying discriminant analysis to a wide range of parameters characterizing a large sample of active regions, we have determined the most common characteristics of a flare-imminent active region's photospheric magnetic field. The results indicate that about a half-dozen properties are important in allowing an active region to be flare productive. Many of these quantities are strongly correlated; physically, large active regions as measured by the total flux also tend to have large vertical currents, significant excess energy, and significant current helicity. Despite the correlations, the top-performing variable for larger flares is the total excess photospheric magnetic energy. Thus, it is not necessary to measure all the correlated quantities in order to distinguish which active regions will be flare producing. However, when modeling active regions, it appears necessary to construct regions that do present all of these (correlated) characteristics in order to best represent the conditions typically found on the Sun.

Interestingly, most of the best variables here are totals of various quantities over the entire region, complemented by measures of the shear, particularly along neutral lines. By including the higher moments of the various distributions, we investigated whether localized areas of, for example, strong vertical current density are important for an active region to flare. The conclusion appears to be that the global properties of the region have more bearing on the flare productivity of the region, while localized variations are not uniquely flare related.

This is surprising, given the association between δ-spots and flaring (e.g., McIntosh 1990). Although merely the presence of a δ-configuration does not necessarily imply the presence of nonpotential field, we find indirect indications that the presence of a δ-spot has a weak association with flaring. A δ-spot contains strong horizontal gradients in the magnetic field, and at least one parameterization of field gradients is present in both Tables 3 and 5. A highly nonpotential δ-spot is likely to have a strongly sheared neutral line, so the parameterizations of neutral line shear, also present in Tables 3 and 5, may also be related to the presence of a δ-spot.

It may also be that the higher order moments are not being recovered well. Particularly for a distribution with a long tail, a large number of points are needed to determine the skew and kurtosis. Even though all the magnetograms considered here have at least 64 pixels with well-measured fields, this may be insufficient to accurately represent localized areas of strong gradients and/or highly nonpotential magnetic field. If the higher order moments are well determined, their absence in the lists of well-performing variables suggests that small patches, which differ from the active region as a whole, do not play an important role in flare production.

It can be seen in Table 4 that even the best discriminant functions do not greatly improve upon the success rate obtained from predicting that nothing will ever flare. That is, over 70% of the data are flare quiet at the C1.0 level, so even with no information about the active region, a 70.38% success rate can still be attained. In comparison, the best-performing discriminant functions of three or more variables only improve the rate of correct classification by about 10%. This is quite a modest improvement, indicating that no variables make a strong distinction between the two states. In part, this is because the majority of active regions are flare quiet during any given 24 hr period, so even though the means of the flaring and flare-quiet samples are quite distinct for many parameters, there is still significant overlap between the populations. The situation is even more extreme when the analysis considers only flares of class M1.0 and larger: the almost 93% success rate of the best discriminant function is only a few percent better than that obtained using a default prediction that no large flares will occur.

Our statistical results are similar to the success rates of Falconer et al. (2006), notwithstanding that their "events" are coronal mass ejections rather than flares, considering that the sample used for that study contained 67% event-quiet regions. Their highest success rates are $\sim$75%, albeit for a much smaller sample, confirming

that the state of the photospheric magnetic field at any particular time does not have a strong influence on the occurrence of energetic events. The approaches to analyzing the data are significantly different, but, as shown in Appendix B, even using nonparametric techniques to estimate the probability density functions is likely to make only small improvements.

Based on our earlier studies, it may be possible to better distinguish a flare-imminent active region by including either the evolution of the photospheric field (Paper II) or the coronal magnetic field (Paper III). The first of these is consistent with the results of Schrijver et al. (2005), who found that the rate of flaring is significantly higher for rapidly evolving active regions. Both of these are supported by various models for the initiation of energetic events, as some require particular coronal magnetic topology (e.g., a coronal null point; Antiochos 1998; Antiochos et al. 1999) or evolution of the photospheric field (e.g., converging flows at polarity inversion lines; Linker et al. 2001; Amari et al. 2003). It may be that including other approaches for characterizing the photospheric magnetic field, such as its fractal dimension or power spectrum of the spatial scales present (Abramenko 2005; McAteer et al. 2005), may significantly improve the success rates. Indeed, simply using observations of the chromospheric magnetic field, which is believed to be force free (Metcalf et al. 2005), may improve the results. However, our results suggest that the state of the photospheric magnetic field at any single time has limited bearing on the occurrence of solar flares.

The work herein was carried out at the Colorado Research Associates Division of NorthWest Research Associates with data from the University of Hawai'i Mees Solar Observatory. We thank E. Schumer for work on an early version of the database and Thomas R. Detman from NOAA/Space Environment Center for the suggestion of using nonparametric techniques. We also thank the referee for helpful suggestions and especially for pointing out our oversight of the Abramenko et al. (1996) paper. Funding from the Air Force Office of Scientific Research is gratefully acknowledged under contract F49620-03-C-0019. This project is dedicated to Barry J. LaBonte.

APPENDIX A

THE IVM QUICK-LOOK DATA

The present analysis represents the first use of the Imaging Vector Magnetograph (IVM) "survey" data in its quick-look form. The quick-look data reduction differs from the "full" data reduction in that it uses a fairly rudimentary flat-fielding approach, takes no account of scattered or parasitic light, and no correction is attempted for seeing variations that occur during the data acquisition.

The inversion from polarization spectra to a magnetic flux map in the image plane (B_l, B_t, ϕ) is performed using a "wavelet analysis," with a transform on each Stokes (I, Q, U, V) spectral profile used to locate the position and amplitude of the components (D. Mickey 2006, private communication). The Paul wavelet is used, since its real component is similar to (Q, U) profiles, and its imaginary part matches the V-profile. The magnetic field values are obtained by multiplying the polarization parameters by a constant, as with "magnetograph" type inversions, and thus suffer from saturation at large fields; additionally, no accounting for magneto-optical effects is included.

The argument made here is that whatever saturation or other errors are systematically present will be present for all regions whether "flaring" or not. If umbral regions saturate and sunspots never achieve the strong fields expected, the estimate of (for example) total magnetic flux will be systematically underestimated as will the spatial gradients of the field distribution. However, with the use of moment analysis of the distributions of derived variables and the discriminant analysis to examine whether there are differences between the two populations, the relevant question to ask is not whether a region had exactly 6×10^{32} Mx of magnetic flux, but rather whether the data can distinguish between two regions that had 6×10^{32} Mx and 3×10^{32} Mx, respectively. The answer to the latter is "yes" to within standard uncertainties, since all data are being treated in the same manner.

We show a comparison between a particular quick-look magnetogram and a full reduction of the same raw data exactly following that used to prepare the time series data for Papers I–III. The region is NOAA AR 09026, a medium-sized active region observed near disk center on 2000 June 05 at 16:30 UT. Figure 4 shows scatter plots for the two different reductions. The data shown are all 2 σ or above, both from plage (*grey*) and sunspots (*black*). The expected saturation is evident (the field strengths from the quick look never achieve that from the full data reduction), the transverse field strength has significant scatter, and there can be significant azimuthal angle differences between the two. This example is fairly clear and typical of the differences found, although of course wide variations do exist. The differences in azimuthal angle are perhaps the most critical, and we point out that in sunspot regions, the vast majority of the points have a difference of less than 20°, or within roughly 2 σ of the normal azimuthal uncertainty.

While any one point in any particular quick-look magnetogram is undoubtedly disputable, we present here a quantitative comparison of the robustness of these data for use in a statistical analysis of solar active regions. We acknowledge that (for example) the possible prevalence of magneto-optical effects may contribute to the correlation between total magnetic flux and (for example) total vertical current. We look forward to a similar-sized database of fully reduced vector magnetograms from future programs and space missions.

APPENDIX B

PROBABILITY DENSITY FUNCTIONS AND THE POSSIBLE ADVANTAGES OF NONPARAMETRIC APPROACHES

To this point in the analysis, the distributions of all the parameters have been assumed to be normal (Gaussian), and the covariance matrices of the two populations have been assumed equal. Clearly, these assumptions cannot be correct for all the parameters considered, although in some cases, they may be a reasonable approximation. With the small samples considered in Papers II and III, it was unreasonable to consider other approximations for the probability distribution functions. However, with the present sample size, two

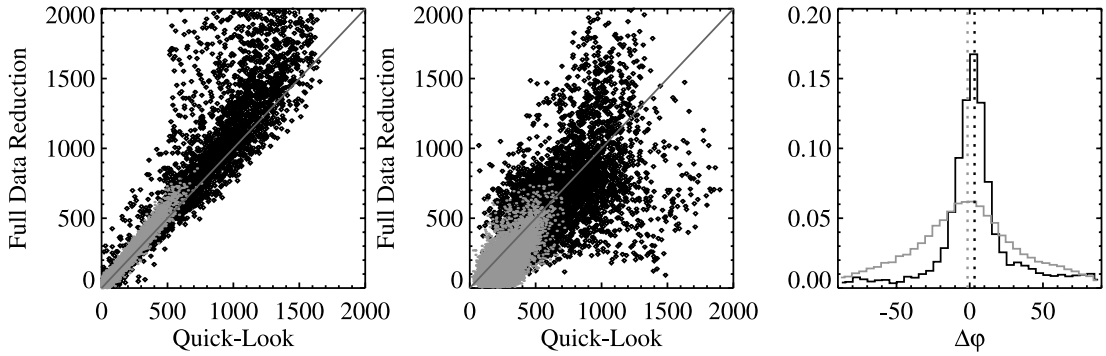

FIG. 4.—Comparison of quick-look and full data reduction for NOAA AR 09026 2000 June 05 obtained at 16:30 UT. *Left*: Longitudinal magnetic field; *middle*: field strength transverse to the line of sight; *right*: angular difference of the azimuthal angle between the two, modulo 90°. Points from sunspots (*black*) and plage area (*grey*) are shown in all three plots, as is the $x = y$ line for the field strengths (*left and middle plots*) and the means of the angular differences (*vertical dashed lines, right plot*) for both sunspots (*black*) and plage areas (*grey*).

ways of relaxing the assumption of normal distributions with equal covariance matrices are investigated. First, the covariance matrix of each population is estimated from its sample, independent of the other sample. Second, a very simple way of approximating the probability distribution function from the sample using a nonparametric technique is considered. Each of these approaches can, for a single-variable example, improve the performance of the discriminant analysis.

B1. UNEQUAL COVARIANCE MATRICES AND THE QUADRATIC DISCRIMINANT

Assume that each of the two samples (flare and nonflaring) comes from a population with a multivariate normal distribution of p variables with mean $\boldsymbol{\mu}^{(i)}$ and covariance matrix $\boldsymbol{\Sigma}^{(i)}$. Using the sample means $\bar{\boldsymbol{x}}^{(i)}$ and covariance matrices $\mathbf{C}^{(i)}$ to estimate the population means and covariance matrices, the probability density functions are estimated as

$$\hat{f}_i(\boldsymbol{x}) = \frac{\sqrt{|\mathbf{C}^{(i)-1}|}}{(2\pi)^p} e^{-(1/2)(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(i)})\mathbf{C}^{(i)-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(i)})}. \tag{B1}$$

Further assume that the a priori probability of membership in the populations is proportional to the sample size. The boundary between predicting a region to flare and predicting it to be flare quiet occurs when the probabilities of flaring and not flaring are equal, so

$$\begin{gathered} n_1 \hat{f}_1(\boldsymbol{x}) = n_2 \hat{f}_2(\boldsymbol{x}), \\ n_1 \sqrt{|\mathbf{C}^{(1)-1}|} e^{-(1/2)(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(1)})\mathbf{C}^{(1)-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(1)})} = n_2 \sqrt{|\mathbf{C}^{(2)-1}|} e^{-(1/2)(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(2)})\mathbf{C}^{(2)-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(2)})}. \end{gathered} \tag{B2}$$

Taking the logarithm of both sides leads to

$$\boldsymbol{x}\left(\mathbf{C}^{(2)-1} - \mathbf{C}^{(1)-1}\right)\boldsymbol{x} - 2\left(\bar{\boldsymbol{x}}^{(2)}\mathbf{C}^{(2)-1} - \bar{\boldsymbol{x}}^{(1)}\mathbf{C}^{(1)-1}\right)\boldsymbol{x} = \bar{\boldsymbol{x}}^{(1)}\mathbf{C}^{(1)-1}\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}\mathbf{C}^{(2)-1}\bar{\boldsymbol{x}}^{(2)} - \ln\left(\frac{n_1^2|\mathbf{C}^{(1)-1}|}{n_2^2|\mathbf{C}^{(2)-1}|}\right), \tag{B3}$$

which is quadratic in $\boldsymbol{x}$, but for $\mathbf{C}^{(1)} = \mathbf{C}^{(2)}$, reduces to the linear expression given in Appendix A of Paper II plus the additional term $\ln(n_1/n_2)$, which comes from assuming the a priori probabilities are proportional to the sample sizes; in Paper II, we assumed equal a priori probabilities because of our selective sample.

Figure 5 shows the linear and quadratic discriminants for the total flux $\Phi_{\rm tot}$ plus the probability density estimate given by normal distributions. Clearly neither probability density is truly normal, but the density for nonflaring regions appears more sharply peaked (at a smaller value of $\Phi_{\rm tot}$) in comparison to the flaring probability density, indicating clear differences in both the means and the covariance matrices of the flaring and nonflaring populations. The linear discriminant is only able to capture the difference in the means, while the quadratic discriminant is able to improve upon the linear discriminant by recovering the greater width of the flaring density but is still not a particularly good representation of the true density.

In Table 7, the best linear and quadratic single-variable discriminant functions are compared based on the classification table error rate. Generally, the quadratic discriminants show a slight improvement compared to the linear discriminants, with all of the same variables being present in the top 10, in slightly different order. In most, but not all cases, the rate of correct classification for the quadratic discriminant is larger than for the linear discriminant.

B2. NONPARAMETRIC ESTIMATES OF THE PROBABILITY DENSITY FUNCTION

If the probability density functions for the variable(s) under consideration were known exactly, the optimal discriminant function could be constructed by predicting a flare wherever the probability density for flaring regions exceeds the probability density for

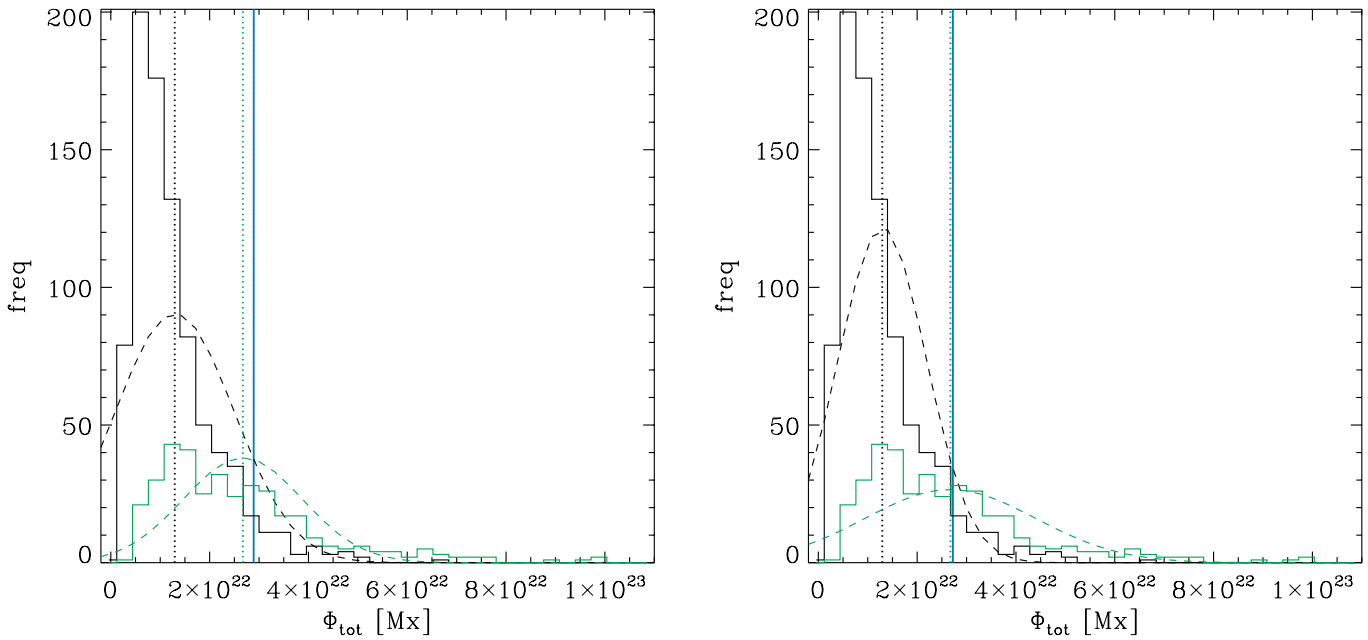

FIG. 5.—Comparison of linear and quadratic discriminant functions for the variable $\Phi_{\rm tot}$, for a C1.0 threshold, in the same format as Fig. 3. *Left*: Equal covariance matrices (linear discriminant) fit is unable to capture the difference in width of the flaring and nonflaring distributions. *Right*: Unequal covariance matrices (quadratic discriminant) fit has a larger spread for the flaring distribution, but is still not a particularly good fit to the true probability density.

flare-quiet regions, adjusted for the a priori probability of membership in each population. That is, predict a flare wherever $n_f f_f(x) \geq n_n f_n(x)$, where the probability of a measurement falling between x_a and x_b is given by

$$P(x_a < x < x_b) = \int_{x_a}^{x_b} f(x)\,dx. \tag{B4}$$

In previous sections, it was assumed that f took the form of a Gaussian distribution, and the parameters describing the Gaussian distribution were estimated from the data. However, the probability density can also be estimated without making any assumption about its functional form as was done in the parametric approach using the Gaussian fit. The kernel method (e.g., Silverman 1986) for estimating the probability density is demonstrated here as a simple and straightforward approach to density estimation.

In the kernel method, the probability density is estimated by summing over the contribution from each data point, weighted by a given kernel function. Given the kernel $K(t)$ and measurements $\{x_1, \cdots, x_n\}$, the probability density is estimated as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(t), \tag{B5}$$

where $t = (x - x_i)/h$ and h is a "smoothing parameter" that must also be selected based on the data. A large value of h will result in an over-smoothed probability density, in which real structure may be hidden, while a small value of h will result in an under-smoothed probability density, in which small-scale structure may be an artifact of the smoothing.

TABLE 7
10 BEST SINGLE-VARIABLE CLASSIFICATION TABLE DISCRIMINANT ANALYSIS RESULTS

LINEAR DISCRIMINANT		QUADRATIC DISCRIMINANT		NONPARAMETRIC DISCRIMINANT	
Variable	Rate	Variable	Rate	Variable	Rate
$\Phi_{\rm tot}$	0.7723	$I_{\rm tot}$	0.7772	E_e	0.7871
$I_{\rm tot}$	0.7682	E_e	0.7756	$I^h_{\rm tot}$	0.7814
$H_c^{\rm tot}$	0.7649	$\Phi_{\rm tot}$	0.7706	$I_{\rm tot}$	0.7772
$\sigma(\Psi_{\rm NL})$	0.7616	$I^h_{\rm tot}$	0.7706	$H_c^{\rm tot}$	0.7756
$I^h_{\rm tot}$	0.7607	$H_c^{\rm tot}$	0.7673	$\Phi_{\rm tot}$	0.7739
$\lvert I^B_{\rm net}\rvert$	0.7484	$\lvert I^B_{\rm net}\rvert$	0.7632	$\sigma(\Psi_{\rm NL})$	0.7673
$\sigma(\rho_e)$	0.7484	$\sigma(\Psi_{\rm NL})$	0.7607	$\lvert H_c^{\rm net}\rvert$	0.7649
$\lvert H_c^{\rm net}\rvert$	0.7450	$\lvert H_c^{\rm net}\rvert$	0.7566	$L(\Psi_{\rm NL} > 45^\circ)$	0.7607
E_e	0.7434	$L(\Psi_{\rm NL} > 45^\circ)$	0.7500	$L(\Psi_{\rm NL} > 80^\circ)$	0.7599
$L(\Psi_{\rm NL} > 45^\circ)$	0.7393	$\sigma(\rho_e)$	0.7442	$\lvert I^B_{\rm net}\rvert$	0.7566

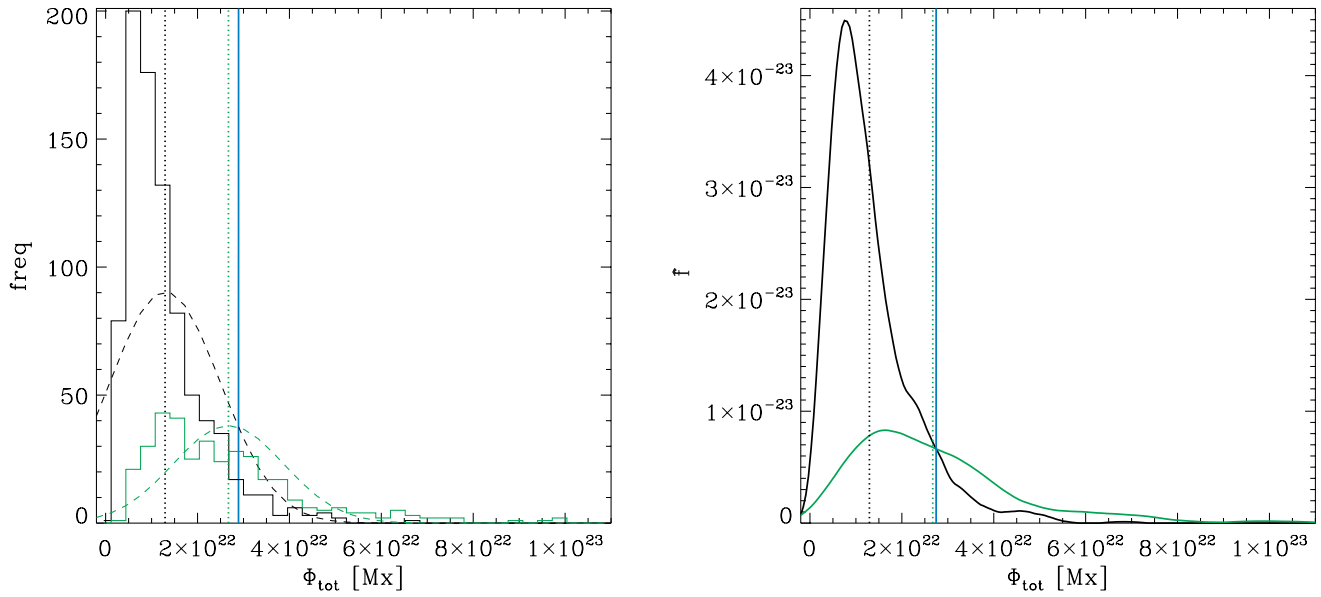

FIG. 6.—Comparison of parametric and nonparametric discriminant functions for the variable Φ_{tot}, in the same format as Fig. 3. *Left:* Parametric representation using Gaussian distributions whose parameters are determined from the sample data, and a nonparametric estimate of the probability density using histograms with optimal bin size (see also Figs. 3 and 5). *Right:* Nonparametric representation of the probability density estimates using the Epanechnikov kernel. The discriminant boundary shown (*blue line*) corresponds to the linear discriminant (Gaussian distributions; *left*) and the nonparametric discriminant boundary (for the Epanechnikov kernel; *right*). Although a different nonparametric discriminant boundary could be calculated from the histograms, the histograms are in fact only included for familiarity. For the Epanechnikov kernel, the bump in the nonflaring density estimate at $\Phi_{tot} \approx 4.5 \times 10^{22}$ Mx followed by the dip at $\Phi_{tot} \approx 6 \times 10^{22}$ Mx are likely examples of the under-smoothing of the tails of the distributions, while the peaks are likely to be over smoothed.

The simplest and best-known nonparametric density estimate is a histogram, which is a convenient way of inspecting data, but it is not the optimal way to estimate the probability density, even when the bin size is selected to optimize the smoothing, as was done here. The most efficient kernel for univariate data is the Epanechnikov kernel (Silverman 1986):

$$K_E(t) = \begin{cases} \dfrac{3}{4\sqrt{5}}\left(1 - \dfrac{t^2}{5}\right), & |t| < 5, \\ 0, & \text{otherwise.} \end{cases} \tag{B6}$$

This kernel is used for all the results presented here, with a smoothing parameter based on the optimum value for the Epanechnikov kernel applied to a normal distribution:

$$h_{opt} \approx \sigma n^{-1/5}, \tag{B7}$$

where σ is the population standard deviation.

The kernel method typically does not work well for distributions with long tails, as either the peak of the distribution will be over smoothed in comparison with the tail, or the tail will be under smoothed in comparison with the peak. Such issues can be dealt with by, for example, using an adaptive kernel method in which the amount of smoothing varies based on the local density estimate. Also, the Epanechnikov kernel does not guarantee that the probability density vanishes for $x < 0$, as must be the case for some variables, such as the total flux. This also can be addressed by, for example, using a kernel that is antisymmetric about 0, but this introduces other difficulties, as the resulting density is no longer normalized to 1. Thus, the nonparametric approach is demonstrated here using only the Epanechnikov kernel with constant h, while noting its limitations.

Figure 6 shows the estimated probability densities for the total flux. Note that neither of the density estimates goes to zero at $\Phi_{tot} = 0$, as should be the case for the reason described above, and that there is a distinct bump in the nonflaring density estimate at $\Phi_{tot} \approx 4.5 \times 10^{22}$ Mx followed by a dip at $\Phi_{tot} \approx 6 \times 10^{22}$ Mx, which is likely due to under-smoothing of the tail of the distribution.

The error rate for a nonparametric discriminant function can be estimated in the same ways as are done for a parametric discriminant function. The classification table is constructed by evaluating the density estimates at each sample data point, and the point is classified as belonging to the group with the larger density estimate. For Φ_{tot}, the result is a correct classification rate of 0.7739, which is a very slight improvement over the linear discriminant. The $n - 1$ error rate is similarly constructed by estimating the probability density at each data point while excluding the data point under consideration. The result is a correct rate of 0.7723. Table 7 lists the top 10 single variable nonparametric discriminants. As in the quadratic discriminant case, most of the same variables appear as the best performers, but in different order, and generally with slight improvements to the classification rate. Note, however, that a few variables, like E_e, have shown a dramatic improvement in comparison with the linear discriminant. These are variables whose distributions are far from Gaussian; in the case of E_e, squaring the field results in a distribution with a very long one-sided tail.

REFERENCES

Abramenko, V. I. 2005, ApJ, 629, 1141
Abramenko, V. I., Wang, T., & Yurchishin, V. B. 1996, Sol. Phys., 168, 75
Amari, T., Luciani, J. F., Aly, J. J., Mikic, Z., & Linker, J. 2003, ApJ, 585, 1073
Anderson, T. W. 1984, An Introduction to Multivariate Statistical Analysis (New York: Wiley)
Antiochos, S. K. 1998, ApJ, 502, L181
Antiochos, S. K., DeVore, C. R., & Klimchuk, J. A. 1999, ApJ, 510, 485
Bao, S., Zhang, H., Ai, G., & Zhang, M. 1999, A&AS, 139, 311
Barnes, G., & Leka, K. D. 2006, ApJ, 646, 1303 (Paper III)
Barnes, G., Longcope, D. W., & Leka, K. D. 2005, ApJ, 629, 561
Bornmann, P. L., & Shaw, D. 1994, Sol. Phys., 150, 127
Canfield, R. C., Hudson, H. S., & McKenzie, D. E. 1999, Geophys. Res. Lett., 26, 627
Canfield, R. C., et al. 1993, ApJ, 411, 362
Falconer, D. A., Moore, R. L., & Gary, G. A. 2003, J. Geophys. Res., 108, 11
———. 2006, ApJ, 644, 1258
Gallagher, P., Moon, Y.-J., & Wang, H. 2002, Sol. Phys., 209, 171
Hagyard, M. J., Smith, J. B. J., Teuber, D., & West, E. A. 1984, Sol. Phys., 91, 115
Hagyard, M. J., Venkatakrishnan, P., & Smith, J. B. J. 1990, ApJS, 73, 159
Hills, M. 1966, J. R. Statis. Soc. B, 28, 1
Kendall, M. G., Stuart, A., & Ord, J. K. 1983, The Advanced Theory of Statistics, Vol. 3 (4th ed.; New York: Macmillan)
Klecka, W. R. 1980, Disciminant Analysis (Beverly Hills: Sage)
Komm, R., Howe, R., Hill, F., González Hernández, I., & Toner, C. 2005, ApJ, 630, 1184
LaBonte, B. 2004, Sol. Phys., 221, 191
LaBonte, B., Mickey, D. L., & Leka, K. D. 1999, Sol. Phys., 189, 1
Leka, K. D., & Barnes, G. 2003a, ApJ, 595, 1277 (Paper I)
———. 2003b, ApJ, 595, 1296 (Paper II)
Leka, K. D., & Skumanich, A. 1999, Sol. Phys., 188, 3
Linker, J. A., Lionello, R., Mikić, Z., & Amari, T. 2001, J. Geophys. Res., 106, 25165
McAteer, R. T. J., Gallagher, P. T., & Ireland, J. 2005, ApJ, 631, 628
McIntosh, P. S. 1990, Sol. Phys., 125, 251
Metcalf, T. R., Leka, K. D., & Mickey, D. L. 2005, ApJ, 623, L53
Metcalf, T. R., et al. 2006, Sol. Phys., 237, 267
Mickey, D. L., Canfield, R., LaBonte, B. J., Leka, K. D., Waterson, M. F., & Weber, H. M. 1996, Sol. Phys., 168, 229
Schrijver, C. J., DeRosa, M. L., Title, A. M., & Metcalf, T. R. 2005, ApJ, 628, 501
Silverman, B. W. 1986, Density Estimation for Statistics and Data Analysis (London: Chapman & Hall)
Wang, J., Shi, Z., Wang, H., & Lü, Y. 1996, ApJ, 456, 861
Wheatland, M. S. 2004, ApJ, 609, 1134
———. 2005, Space Weather, 3, S07003
Zhang, H. 2001, ApJ, 557, L71