The Astrophysical Journal, 688: L107–L110, 2008 December 1

EVALUATING THE PERFORMANCE OF SOLAR FLARE FORECASTING METHODS

G. Barnes and K. D. Leka
Colorado Research Associates Division, NorthWest Research Associates, 3380 Mitchell Lane, Boulder, CO 80301; graham@cora.nwra.com
Received 2008 July 24; accepted 2008 October 7; published 2008 November 3

ABSTRACT

The number of published approaches to solar flare forecasting using photospheric magnetic field observations has proliferated recently, with widely varying claims about how well each works. As different analysis techniques and data sets were used, it is essentially impossible to directly compare the results. A systematic comparison is presented here using three parameters based on the published literature that characterize the photospheric magnetic field itself, plus one that characterizes the coronal magnetic topology. Forecasts based on the statistical method of discriminant analysis are made for each of these parameters, and their ability to predict major flares is quantified using skill scores. Despite widely varying statements regarding their forecasting utility in the original studies describing these four parameters, there is no clear distinction in their performance here, thus demonstrating the importance of using standard verification statistics.

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

1. INTRODUCTION

Recently, several new parameters derived from observations of the photospheric magnetic field have been proposed as useful in determining the flare productivity of an active region (Georgoulis & Rust 2007; Schrijver 2007). It is claimed that one "is an efficient flare-forecasting criterion" (Georgoulis & Rust 2007, hereafter GR07), while the other "can therefore be used effectively for flare forecasting" (Schrijver 2007, hereafter S07). In comparison, a recent study by Leka & Barnes (2007) that included a large number of parameters characterizing the photospheric magnetic field concluded that "the state of the photospheric magnetic field at any given time has limited bearing on whether that region will be flare productive." In each case, the qualitative statements are based on different ways of assessing the performance of the parameter(s) that are derived from different databases, in some cases using a different definition of event. It is therefore not possible to make a quantitative comparison of these results from the literature.

To make such a comparison, we have computed parameters based on those proposed by GR07 and S07 for the database of vector magnetograms described in Leka & Barnes (2007) and made forecasts using a statistical approach based on discriminant analysis. With this consistent approach, we present here a careful comparison of the performance of four illustrative parameters, including two used in Leka & Barnes (2007). In addition to intercomparing the success rates, skill scores were constructed from the forecasts. A skill score gives a normalized measure of how well a forecasting technique does in comparison to making uniform (climatological) forecasts, for which no information about any particular active region is needed. These validation statistics are in common use in the meteorology community, have also been used by Wheatland (2005) for an approach to flare forecasting based on flare persistence, and are published by the National Weather Service/Space Weather Prediction Center (SWPC, formerly NOAA/SEC) for their flare forecasts.[1] Although care must still be taken in interpreting the results, the use of validation statistics provides a quantitative way of comparing the performance of approaches to flare forecasting that use different data sources.

The performance of the four representative parameters with respect to major flare daily forecasting was found to be quite similar; indeed all four are moderately to strongly correlated with each other (linear correlation coefficients $0.78 \leq r \leq 0.93$). At best there is modest improvement over making uniform (climatological) forecasts. However, the modest improvements are comparable to those of flare forecasting approaches based on completely different data sources. Thus, observations of the photospheric magnetic field appear to yield just as much information about whether a solar flare is imminent as does past flaring history (e.g., Wheatland 2004) or white-light observations and associated climatology (e.g., McIntosh 1990; Bornmann & Shaw 1994).

2. DATA AND EVENT DEFINITION

To test their performance for flare forecasting, parameters were calculated for the magnetograms in the database described in Leka & Barnes (2007). The database consists of 1212 single "quick look" vector magnetograms taken by the Imaging Vector Magnetograph (IVM; Mickey et al. 1996; LaBonte et al. 1999) during routine observations from 2001 to 2004. All numbered active regions are included; the only selection criteria that were imposed are that the center of the field of view not be farther from disk center than $\mu = \cos\theta = 0.5$, that at least 64 data points exist for which both the line-of-sight field and the transverse field were greater than the 2 σ level, and that the data are free of visible defects. Additional details can be found in Leka & Barnes (2007), including a discussion in Appendix A of the limitations of the quick-look data reduction; the data are available on the Web.[2]

A region is classified as producing a (major) flare if the event logs for the *Geostationary Operational Environmental Satellite* (*GOES*) available through the National Geophysical Data Center[3] record at least one event with peak soft X-ray flux greater than or equal to 1.0×10^{-5} W m^{-2}, corresponding to an M- or X-class flare, in the 24 hr following the time of the magnetogram. This definition matches that of S07, but differs from that given in GR07, who considered events that occurred prior to, during, and after the 12 hr interval over which magnetogram data were averaged to construct their parameter. The goal here is a systematic comparison of flare forecasting parameters; thus we employ an event definition consistent with that goal. With this definition of event, there are 111 flaring regions in the database.

[1] See http://www.swpc.noaa.gov/forecast_verification/.

[2] See http://www.cora.nwra.com/~ivm/IVM_SurveyData/.

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

3. THE PARAMETERS

The parameters considered consist of the total unsigned magnetic flux, plus one each from Leka & Barnes (2007) and S07 characterizing the photospheric magnetic field, and one from GR07 characterizing the coronal connectivity. The total flux is given by

$$\Phi_{\rm tot} = \int d^2x \, |B_z|, \tag{1}$$

where the integral is approximated by a sum over the field of view. It is a simple measure of the size of an active region, and is often viewed as a standard against which to compare other forecasting parameters.

3.1. *From Leka & Barnes (2003a)*

One of the best performing parameters considered by Leka & Barnes (2007) for predicting the occurrence of large flares is the "total excess energy"

$$E_e = \int d^2x \, (\boldsymbol{B} - \boldsymbol{B}_p)^2, \tag{2}$$

where $\boldsymbol{B}_p$ is the potential field, and the integral is again approximated by a sum over the field of view. This integral of the deviation from the potential state is taken as a proxy for the magnetic free energy (Leka & Barnes 2003a); thus one expects that major flares will only come from regions with a large value of E_e.

3.2. *From Schrijver (2007)*

The next parameter considered here was proposed by S07, and is a measure of the amount of magnetic flux close to high-gradient polarity-separation lines. It was interpreted by S07 as a proxy for the emergence of current-carrying flux. To compute this parameter R, bitmaps of the magnetograms where the positive or negative flux density exceeds 150 Mx cm^{-2} were dilated with a kernel of 3×3 pixels^2. The high-gradient polarity-separation lines were defined to be the areas where the bitmaps overlap. Then, the bitmap of the high-gradient polarity-separation lines was convolved with a Gaussian to obtain a weighting map. Finally, the unsigned flux close to these areas was determined by multiplying the weighting map by the unsigned line-of-sight field to obtain the parameter R.

This parameter was calculated by S07 for MDI magnetograms that have a pixel size of $2'' \times 2''$. In order to approximately match this spatial scale, the IVM magnetograms used here were spatially binned, resulting in $2.2'' \times 2.2''$ pixels. Even though the vertical magnetic field is available for the magnetograms used, the line-of-sight component was used to compute the "flux" to match as closely as possible what was done by S07. In addition, the convolution was done with a Gaussian of width 10 (rebinned) pixels, not a physical distance, to again match as closely as possible the calculations of S07.

3.3. *From Georgoulis & Rust (2007)*

The final parameter, as proposed by GR07, is based on a magnetic charge topology (MCT) model (Baum & Bratenahl 1980; Gorbachev & Somov 1988; Lau 1993), in which the contribution to the coronal magnetic field by each concentration of magnetic flux at the photosphere is represented by the field of a magnetic point source. This class of model has the advantage that the coronal magnetic topology becomes particularly simple: with a few special exceptions, each field line must start on a source of one polarity, and end on a source of the opposite polarity. It is therefore straightforward to define a connectivity matrix whose elements, ψ_{ij}, comprise the magnetic flux connecting source i with source j.[4] To determine the point sources, each magnetogram was partitioned following the algorithm used by GR07, first described in Barnes et al. (2005) with a smoothing parameter $h = 0.5$ Mm and a saddle point parameter $B_s = 100$ G.

From the connectivity matrix and the locations of the sources, $\boldsymbol{x}_i$, GR07 define the "effective connected magnetic field,"

$$B_{\rm eff} = \frac{1}{2} \sum_{i \neq j} \frac{\psi_{ij}}{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}, \tag{3}$$

where $\boldsymbol{x}_i$ is the position of source i. This quantity is extremely similar to the parameter

$$\phi_{\rm tot} = \frac{1}{2} \sum_{i \neq j} \frac{\psi_{ij}}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} \tag{4}$$

previously considered by Barnes & Leka (2006). For the database used here, the correlation coefficient between $B_{\rm eff}$ and $\phi_{\rm tot}$ is 0.97, and the forecasting ability of the two parameters is extremely similar; thus it is hard to believe that the additional factor of $|\boldsymbol{x}_i - \boldsymbol{x}_j|^{-1}$ in $B_{\rm eff}$ plays any significant role. Thus, we evaluate only whether our rendition of $B_{\rm eff}$ is a robust parameter for major flare forecasting.

One difference in our computation of $B_{\rm eff}$ is the method used for determining the connectivity matrices. The connectivity matrices computed herein were calculated by tracing field lines and employing the Bayesian estimate described in Barnes et al. (2005). For this analysis, the number of field lines was chosen to give a detection threshold of $\psi_c = 15$ G Mm^2. That is, typically 95% of domains with a flux $\psi_{ij} > \psi_c$ will be found.

GR07 determined a connectivity matrix by using simulated annealing to minimize the function

$$F = \sum_{i=1}^{N_+} \sum_{j=1}^{N_-} \left(\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|}{|x_i| + |x_j|} + \frac{|\Phi_i' + \Phi_j'|}{|\Phi_i'| + |\Phi_j'|} \right), \tag{5}$$

where Φ_i' is the flux of source i in flux "units" and N_+, N_- are the number of positive and negative sources. The physical meaning of the connectivity arrived at by this method is unclear, since the function being minimized depends on the choice of origin of the coordinate system. The resulting connectivity, in general, matches neither the potential field connectivity nor the true coronal connectivity. Since minimizing F selects for the shortest connections available, the parameter $B_{\rm eff}$ constructed from this connectivity could perhaps be viewed as having a similar interpretation to the parameter proposed by S07, measuring the amount of flux close to polarity inversion lines, but in this case with a weighting determined by the connectivity, rather than a Gaussian. Our field line tracing code appears to be at least as fast as the simulated annealing approach, recovers the potential field connectivity thus giving it a physical meaning, and hence was used here.

[4] We shall use the notation of Barnes et al. (2005) which differs from that used by GR07, but the values of the resulting parameter do not change.

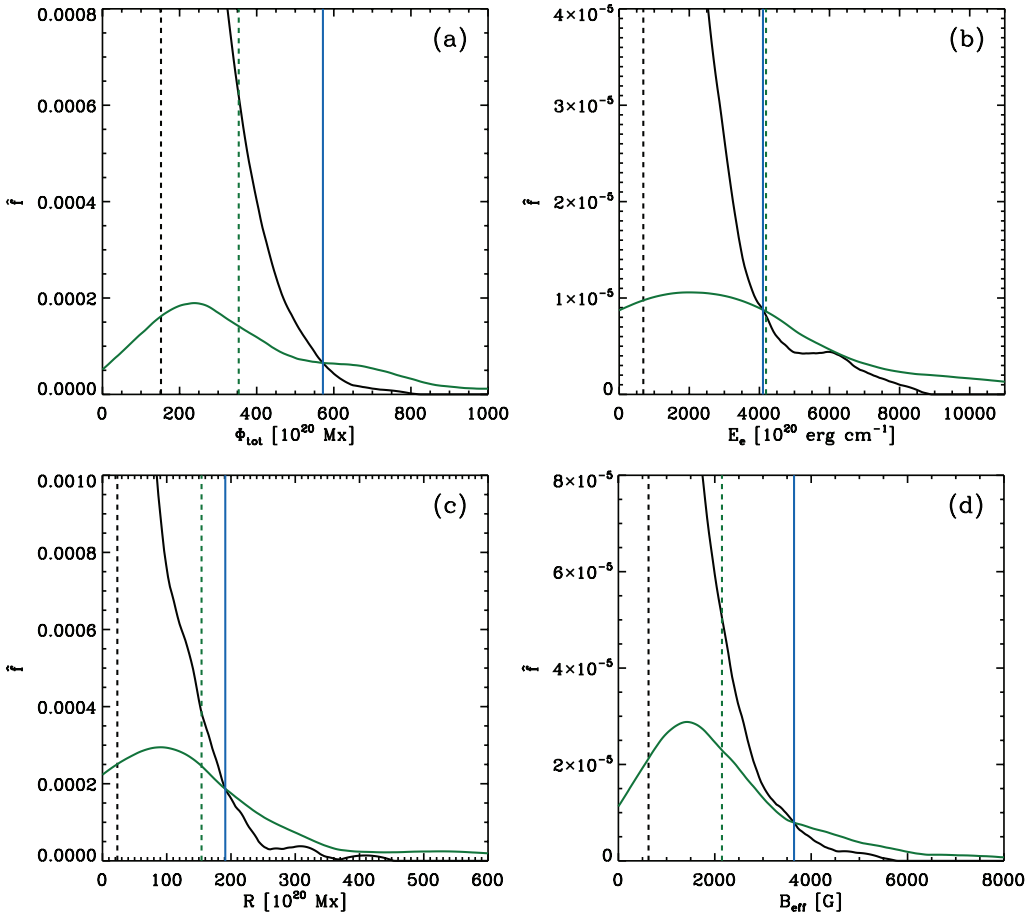

FIG. 1.—Nonparametric density estimates for the flaring (*green*) and nonflaring (*black*) populations for (*a*) Φ_{tot}, (*b*) E_e, (*c*) R, and (*d*) B_{eff}. The discriminant boundary (50% probability forecast) is shown as a vertical blue line, and the sample means are shown as black/green vertical dashed lines. All the parameters exhibit similar behavior, with a tendency for regions with large parameter values to be more likely to produce an event, but no clear separation between the populations.

4. RESULTS FROM DISCRIMINANT ANALYSIS AND PROBABILITY FORECASTS

To compare the forecasting ability of the parameters, the results of discriminant analysis (Kendall et al. 1983; Leka & Barnes 2003b, 2007) were used. In this approach, a nonparametric estimate of the probability density is used to minimize the overall rate of misclassifications by forecasting a region to flare whenever the probability density estimate for flaring regions exceeds the probability density estimate for nonflaring regions. Probability density estimates for the two populations using the Epanechnikov kernel (Silverman 1986), along with the discriminant boundary, are shown in Figure 1. Qualitatively, all the parameters have similar distributions, with significant overlap between the flaring and nonflaring populations. Both the flaring and nonflaring distributions peak at relatively small parameter values, although the flaring population in each case has a longer tail to large parameter values.

TABLE 1
SUCCESS RATES AND SKILL SCORES FOR THE SAMPLE PARAMETERS

Parameter	Success Rate	Heidke Skill Score	Climatological Skill Score
Climatology	0.908	0.000	0.000
Φ_{tot}	0.922	0.153	0.197
E_e	0.916	0.081	0.231
R	0.922	0.144	0.242
B_{eff}	0.913	0.072	0.220

The forecasting ability of a parameter was first evaluated by estimating its success rate, that is, the fraction of correct classifications. An unbiased estimate of the success rate was obtained by removing one point from the data, using the remaining $n - 1$ points to make a prediction about the removed point, and repeating for all n points in the sample (Hills 1966). The performance of the sample parameters, shown in the first column of Table 1, initially looks impressive, with success rates over 90%. However, it is important to realize that simply predicting that no region will ever produce an event can result in quite a high success rate because the majority of active regions do not produce any large flares within a 24 hr period (e.g., McAteer et al. 2005; Leka & Barnes 2007); for the data shown here, 90.8% of the regions produced no flares of at least M class within 24 hr, so a flare forecasting parameter must get more than 90.8% correct for it to add any value.

This drawback to interpreting the success rate when the occurrence of events is very rare has been known in the meteorology community for over a century, and is often illustrated by the study of Finley (1884) (see Murphy 1996 for an overview of the "Finley affair" and the response it provoked). A plethora of ways to quantify the performance of a forecasting method while taking into account the frequency of events have

been discovered and rediscovered over the last century. One is the Heidke skill score (e.g., Wilks 1995), given by

$$\mathrm{SS} = \frac{n_{ff} + n_{qq} - n_q}{n_f}, \tag{6}$$

where n_{ff} is the number of regions that were predicted to flare and did flare, n_{qq} is the number of regions that were predicted to remain flare-quiet and did so, n_f is the number of regions that produced a flare, and n_q is the number of regions that did not produce a flare. This skill score indicates the improvement of the forecasts over always forecasting that no flare will occur. Positive scores indicate better performance, with a maximum score of 1.0 for perfect forecasting, while negative scores indicate worse performance. The Heidke skill scores for the sample parameters are shown in the second column of Table 1. The general results are very similar to evaluations using the success rate, in that there is little variation among parameters, but now the small values of the skill score ($\mathrm{SS} < 0.16 \ll 1$) clearly indicate that there is only slight improvement over forecasting that no flares will occur ($\mathrm{SS} = 0$).

As an alternative to the binary classification of the points presented above, Bayes' theorem can be used to estimate the probability of a flare occurring:

$$P_f(x) = \frac{q_f f_f(x)}{q_f f_f(x) + q_q f_q(x)}, \tag{7}$$

where q_j is the prior probability of belong to population j, estimated as $q_j = n_j/(n_f + n_q)$, $f_j(x)$ is the probability density function, and here $j \in \{f, q\}$. One way to assess the performance of probability forecasts is the climatological skill score (e.g., Murphy & Epstein 1989), defined by

$$\mathrm{SS}(P_f, x) = 1 - \mathrm{MSE}(P_f, x)/\mathrm{MSE}(\langle x \rangle, x) \tag{8}$$

where $\mathrm{MSE}(P_f, x) = \langle (P_f - x)^2 \rangle$ is the mean square error. This skill score indicates the improvement of the forecasts over a constant forecast given by the average observed rate, $\langle x \rangle$. The interpretation of this skill score is similar to the previous one, with a maximum score of 1.0 for perfect forecasting, and 0.0 for a climatological (uniform probability) forecast. This skill score for the sample parameters is shown in the final column of Table 1. The results ($0.20 \lesssim \mathrm{SS} \lesssim 0.24$) are slightly better than for the binary forecasts, but still show only modest improvements over uniform forecasts ($\mathrm{SS} = 0$). Values of this skill score are also quoted by Wheatland (2005) as 0.258 for his approach to forecasting, and 0.262 for the published results from the SWPC. These values are based on a different data set, with a somewhat different definition of event, so care should still be taken in making the comparison. However, using skill scores to account for the differences, it appears that forecasts based on magnetic field observations can perform comparably to other approaches. Barnes et al. (2007) presented a detailed comparison of combinations of parameters characterizing the photospheric magnetic field to the results of Wheatland (2005) and the SWPC.

5. CONCLUSIONS

Using only nonparametric discriminant analysis and the related probability forecasts, the empirical parameters considered here were not found to be robust daily flare predictors. The results for all these parameters are extremely similar: there is substantial overlap in the estimated probability density functions, with the flaring probability density only exceeding the nonflaring probability density in the tail of the distributions.

Although the parameters have not necessarily been calculated in the exact fashion proposed, all have high success rates, as in the initial studies. The highest success rate for any single parameters was 92.2%, yet this is only a slight improvement compared to the success rate of 90.8% obtained by forecasting that no region will ever produce an M or larger flare. Rather than quoting success rates alone, which at the least should be compared to the event rate, we encourage other investigators to make use of standard verification statistics that account for the mean event rate. In this case, skill scores $\mathrm{SS} \lesssim 0.24$ clearly show that the parameters considered here show only modest improvements over uniform or climatological forecasts after accounting for the low mean event rate.

Although the improvements in forecast performance are fairly small for the parameters considered here, the best magnetic parameters already perform comparably to independent approaches for flare forecasting. Improvements may come from treating flares as an example of self-organized criticality (e.g., Bélanger et al. 2007), considering the evolution of the magnetic field, or combining magnetic field observations with independent quantities, such as flaring history. Ideally, comparisons of the performance of existing and new approaches to flare forecasting should be made from the same database, but the use of skill scores can make comparisons based on different data sources more meaningful.

This work was supported by the Air Force Office of Scientific Research under contracts F49620-03-C-0019 and FA9550-06-C-0019, and by the NASA/JSC Space Radiation Analysis Group. We thank Karel Schrijver and the two referees for helpful comments.

REFERENCES

Barnes, G., & Leka, K. D. 2006, ApJ, 646, 1303
Barnes, G., Longcope, D. W., & Leka, K. D. 2005, ApJ, 629, 561
Barnes, G., et al. 2007, Space Weather, 5, S09002
Baum, P. J., & Bratenahl, A. 1980, Sol. Phys., 67, 245
Bélanger, E., Vincent, A., & Charbonneau, P. 2007, Sol. Phys., 245, 141
Bornmann, P. L., & Shaw, D. 1994, Sol. Phys., 150, 127
Finley, J. P. 1884, Am. Meteor. J., 1, 85
Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 661, L109
Gorbachev, V. S., & Somov, B. V. 1988, Sol. Phys., 117, 77
Hills, M. 1966, J. R. Statis. Soc. B, 28, 1
Kendall, M., Stuart, A., & Ord, J. K. 1983, The Advanced Theory of Statistics, 4th ed., Vol. 3 (New York: Macmillan)
LaBonte, B., Mickey, D. L., & Leka, K. D. 1999, Sol. Phys., 189, 1
Lau, Y.-T. 1993, Sol. Phys., 148, 301
Leka, K. D., & Barnes, G. 2003a, ApJ, 595, 1277
Leka, K. D., & Barnes, G. 2003b, ApJ, 595, 1296
———. 2007, ApJ, 656, 1173
McAteer, R. T. J., Gallagher, P. T., & Ireland, J. 2005, ApJ, 631, 628
McIntosh, P. S. 1990, Sol. Phys., 125, 251
Mickey, D. L., et al. 1996, Sol. Phys., 168, 229
Murphy, A. H. 1996, Weather Forecasting, 11, 3
Murphy, A. H., & Epstein, E. S. 1989, Mon. Weather Rev., 117, 572
Schrijver, C. J. 2007, ApJ, 655, L117
Silverman, B. W. 1986, Density Estimation for Statistics and Data Analysis (London: Chapman and Hall)
Wheatland, M. S. 2004, ApJ, 609, 1134
———. 2005, Space Weather, 3, S07003
Wilks, D. S. 1995, Statistical Methods in the Atmospheric Sciences (San Diego: Academic Press)