Solar Phys (2009) 254: 101–125
DOI 10.1007/s11207-008-9288-3

Statistical Assessment of Photospheric Magnetic Features in Imminent Solar Flare Predictions

Hui Song · Changyi Tan · Ju Jing · Haimin Wang · Vasyl Yurchyshyn · Valentyna Abramenko

Received: 5 April 2007 / Accepted: 29 October 2008 / Published online: 28 November 2008

Abstract In this study we use the ordinal logistic regression method to establish a prediction model, which estimates the probability for each solar active region to produce X-, M-, or C-class flares during the next 1-day time period. The three predictive parameters are (1) the total unsigned magnetic flux T_{flux}, which is a measure of an active region's size, (2) the length of the strong-gradient neutral line L_{gnl}, which describes the global nonpotentiality of an active region, and (3) the total magnetic dissipation E_{diss}, which is another proxy of an active region's nonpotentiality. These parameters are all derived from SOHO MDI magnetograms. The ordinal response variable is the different level of solar flare magnitude. By analyzing 174 active regions, L_{gnl} is proven to be the most powerful predictor, if only one predictor is chosen. Compared with the current prediction methods used by the Solar Monitor at the Solar Data Analysis Center (SDAC) and NOAA's Space Weather Prediction Center (SWPC), the ordinal logistic model using L_{gnl}, T_{flux}, and E_{diss} as predictors demonstrated its automatic functionality, simplicity, and fairly high prediction accuracy. To our knowledge, this is the first time the ordinal logistic regression model has been used in solar physics to predict solar flares.

Keywords Sun: Activity · Sun: Flares · Sun: Magnetic fields · Magnetic observables

H. Song · C. Tan (✉) · J. Jing · H. Wang
Space Weather Research Laboratory, Center for Solar-Terrestrial Research, New Jersey Institute of Technology, Newark, NJ 07102, USA
e-mail: ct45@njit.edu

H. Wang
e-mail: haimin@flare.njit.edu

V. Yurchyshyn · V. Abramenko
Big Bear Solar Observatory, New Jersey Institute of Technology, 40386 North Shore Lane, Big Bear City, CA 92314, USA

1. Introduction

Over the past decades, humankind has become more and more dependent on space systems, satellite-based services, and various ground-based facilities. All these technologies are influenced by Sun – Earth interactions. Therefore, one of the primary objectives in space weather research is to predict the occurrence of solar flares and coronal mass ejections (CMEs), which are believed to be the major causes of geomagnetic disturbances (*e.g.*, Brueckner *et al.*, 1998; Cane, Richardson, and St. Cyr, 2000; Gopalswamy *et al.*, 2000; Webb *et al.*, 2000; Wang *et al.*, 2002; Zhang *et al.*, 2003).

It has long been known that solar flares tend to occur along magnetic polarity inversion lines where the magnetic field lines are often highly sheared, with the transverse field directed nearly parallel to the polarity inversion line (Svestka, 1976; Hagyard *et al.*, 1984; Sawyer, Warwick, and Dennett, 1986). Canfield, Hudson, and McKenzie (1999) showed that CMEs also tend to arise in connection with active regions (ARs) exhibiting strong sheared and/or twisted coronal loops called sigmoids. The twisting, tangling, and shearing of magnetic loops lead to magnetic topological complexities and build up a stressed flux system (and excess energy). Subsequent destabilizing events such as local emergence of new magnetic flux from below the photosphere or changes in magnetic connectivity owing to magnetic field reorganization elsewhere on the Sun may result in the release of energy (Hess, 1964; Svestka, 1976; Priest and Forbes, 2000).

To date, various observational studies have explored the connection between photospheric magnetic fields and solar flares, supporting the hypothesis that solar flares are driven by the nonpotentiality of magnetic fields (Moreton and Severny, 1968; Abramenko, Gopasyuk, and Ogir', 1991; Leka *et al.*, 1993; Wang, Xu, and Zhang, 1994; Wang *et al.*, 1996; Tian, Liu, and Wang, 2002; Abramenko, 2005). Through five solar flares, Wang (2006) found there are obvious changes of the magnetic gradient occurring immediately and rapidly following the onset of each flare. Falconer (2001) and Falconer, Moore, and Gary (2003) measured the lengths of strong-sheared and strong-gradient magnetic neutral line segments and found that they are strongly correlated with CME productivity of an active region and both might be prospective predictors. In a study of six large (X5 or larger) flares, Wang *et al.* (2006) reported a positive linear relationship between the magnetic shear and the magnetic gradient and that the latter seems to be a better tool to predict the occurrence of flares and CMEs in an active region. According to Song *et al.* (2006), the length of the strong-gradient neutral line, $L_{\rm gnl}$, was proved to be a viable tool to locate source regions of either CMEs or flares. The definitive flare/CME prediction ability by measuring $L_{\rm gnl}$ is about 75% (55 out of 73 events). Jing *et al.* (2006) analyzed three magnetic parameters – *i*) the mean spatial magnetic field gradient at the strong-gradient magnetic neutral line, $M_{\rm gnl}$; *ii*) the length of a strong-gradient magnetic neutral line, $L_{\rm gnl}$; and *iii*) the total magnetic energy dissipation, $E_{\rm diss}$, which describes magnetic field features of nonpotentiality and turbulence – and found that these parameters have a positive correlation with the overall flare productivity of ARs. Active regions with larger $M_{\rm gnl}$, $L_{\rm gnl}$, and $E_{\rm diss}$ generally show a higher incidence of flaring activity.

The purpose of this study is to find out whether statistical methods that are conceptually simple and algorithmically fast are able to provide a feasible way to evaluate the probability of an active region in producing solar flares. The ordinal logistic regression model satisfies our criteria. The model describes the relationship between an ordered response variable and a set of predictive variables. In our case, the ordered response variable represents four different energy levels of solar flares. We assign numerical values 3, 2, 1, and 0 to represent X-, M-, C-, and B-class flares, respectively. The predictive variables so far include $L_{\rm gnl}$

and $E_{\rm diss}$, which were used in the study by Jing *et al.* (2006), and total unsigned magnetic flux, $T_{\rm flux}$. Mathematically, what the ordinal regression model describes is not the value of the response variable itself but the probability, Prob, that it assumes a certain response value (0, 1, 2, or 3). Thus, in this study, Prob represents the probability of a certain class of flare to occur. Since Prob ranges from 0 to 1, traditional linear regression is inappropriate to predict its value directly.

We will study whether the ordinal logistic regression model is able to predict the occurrence of solar flares in the next 1-day period. The remainder of this paper proceeds as follows. In Section 2 the data sets used to perform the statistical analysis are described. Three magnetic measures are calculated based on the full-disk Michelson Doppler Imager (MDI) magnetograms. In Section 3, the ordinal logistic model is specified and established. The results obtained from the statistical regression model are presented in Section 4, and Section 5 concludes this paper with a discussion of key results.

2. Methods

2.1. Data Collection

Solar activity reports are available online from the U.S. National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC).[1] The reports include detailed information about solar flares, such as the coordinated universal time (UTC) of the beginning, maximum, and end of a flare, the X-ray flux at the flare peak, and the location of the flare, if available. Our study focuses on those flares occurring between 1996 and 2005. Flares were selected according to two criteria: (1) the location of the flare must be accurately indicated in the reports and as close to disk center as possible ($\pm 40^\circ$ in longitude and $\pm 40^\circ$ in latitude), so that extreme projection effects of magnetic fields can be avoided; (2) MDI full-disk magnetograms onboard the *Solar and Heliospheric Observatory* (SOHO) must be available. Additionally, to correct the slight projection effect, we construct the heliographic plane following the method of Gary and Hagyard (1990) under the radial field approximation, under which it is assumed that the measured magnetic fields are radially pointing inward or outward from the Sun. The corrected vertical field is obtained by interpolating fitted values into the heliographic plane. The advantage of SOHO/MDI is that the data can be routinely obtained through the Internet and the observations are free of atmospheric seeing. However, note that the MDI full-disk magnetic field measurements are known to underestimate the actual line-of-sight field component (Berger and Lites, 2003). The flux underestimate in active regions is significant ($\sim$30%) and this has to be corrected for. The flux density underestimate factors derived in Berger and Lites's analysis are roughly 0.65 in plage regions and 0.70 in active regions (umbra and penumbra) for MDI full-disk magnetograms. If there is no saturation, the relationship is almost linear. Therefore, we avoided those saturated MDI magnetograms and employed the simple correction formula

$$B_{\rm zMDI}^{\rm crt} = B_{\rm zMDI}^{\rm raw}/0.7, \tag{1}$$

where $B_{\rm zMDI}^{\rm crt}$ is the corrected MDI magnetogram and $B_{\rm zMDI}^{\rm raw}$ is the underestimated one (Berger and Lites, 2003). To verify this underestimate, we performed an experimental comparison of the magnetograms from MDI and the newly launched *Hinode* SOT/SP (see the

[1] http://www.swpc.noaa.gov/ftpmenu/indices.html.

Appendix). Our experiment shows the difference of each *Hinode* SOT/SP, and the adjusted MDI measurement is smaller than or around 10%; thus this bias would not influence our results much. In the Appendix, it is proven that our corrections of the measurements are valid. In this study, 230 solar flare events (174 ARs) were chosen for analysis.

2.2. Definition of the Predictive and Response Variables

Detailed descriptions of how the photospheric magnetic parameters are calculated from the MDI magnetograms are presented in detail in Jing *et al.* (2006). Thus, we will only briefly list them here:

1. The total unsigned magnetic flux, $T_{\rm flux}$, is a measure of the active region's size.
2. The length of the strong-gradient neutral line, $L_{\rm gnl}$, describes the global nonpotentiality of an active region. The spatial gradient is calculated as

$$\nabla B_z = \left[\left(\frac{\mathrm{d}B_z}{\mathrm{d}x}\right)^2 + \left(\frac{\mathrm{d}B_z}{\mathrm{d}y}\right)^2\right]^{1/2}, \tag{2}$$

   where B_z is the line-of-sight components of the magnetic field measured in the plane (x, y). The gradient threshold in this paper was chosen to be 50 G Mm^{-1} (Falconer, Moore, and Gary, 2003; Song *et al.*, 2006; Jing *et al.*, 2006; Tan *et al.*, 2007).
3. The total magnetic energy dissipation of B_z is $E_{\rm diss} = \int \varepsilon(B_z)\,\mathrm{d}A$, where the integration is done over the entire active region area A and $\varepsilon(B_z)$ is defined according to the following expression (Abramenko *et al.*, 2003):

$$\varepsilon(B_z) = \left(4\left[\left(\frac{\mathrm{d}B_z}{\mathrm{d}x}\right)^2 + \left(\frac{\mathrm{d}B_z}{\mathrm{d}y}\right)^2\right] + 2\left(\frac{\mathrm{d}B_z}{\mathrm{d}x} + \frac{\mathrm{d}B_z}{\mathrm{d}y}\right)^2\right). \tag{3}$$

   Abramenko *et al.* (2003) explained that this measure indicates the dissipative breakdown of the turbulent inertial range and the length scale of flux tubes that are much smaller than 2 – 3 Mm. However, $\varepsilon(B_z)$ can only provide information down to the spatial resolution of the observing instrument. For full-disk MDI measurements the spatial resolution is indeed ∼3 Mm (pixel size of ∼1.98 arc sec, or ∼1.45 Mm). Because the gradient of B_z is also included in $\varepsilon(B_z)$, it could be another proxy of an active region's nonpotentiality.

Flare productivity can be quantitatively described by the flare index. The overall flare productivity of a given active region, which is quantified by weighting the soft X-ray (SXR) flares of X-, M-, C- and B-class as 100, 10, 1, and 0.1, respectively (Antalova, 1996; Abramenko, 2005), is given by

$$F_{\rm idx} = \left(100 \times \sum_{\tau} I_{\rm X} + 10 \times \sum_{\tau} I_{\rm M} + 1 \times \sum_{\tau} I_{\rm C} + 0.1 \times \sum_{\tau} I_{\rm B}\right)/\tau, \tag{4}$$

where τ is the length of time (measured in days) during which an active region is visible on the solar disk and $I_{\rm X}$, $I_{\rm M}$, $I_{\rm C}$, and $I_{\rm B}$ are GOES peak intensities of X-, M-, C-, and B-class flares produced by a given active region for the duration τ. To evaluate the flare production of an active region in the next 1-day time interval, τ is selected to be 1.

As an example, in Figure 1 we present the calculation of these parameters for NOAA AR 9077 on 14 July 2000. The left panel shows the MDI line-of-sight magnetogram of this flare-active region. The overall $F_{\rm idx}$ (accumulated for about 13 days while it passed through the solar disk) is as high as 1256.40×10^{-6} W m^{-2}, equivalent to a specific flare productivity

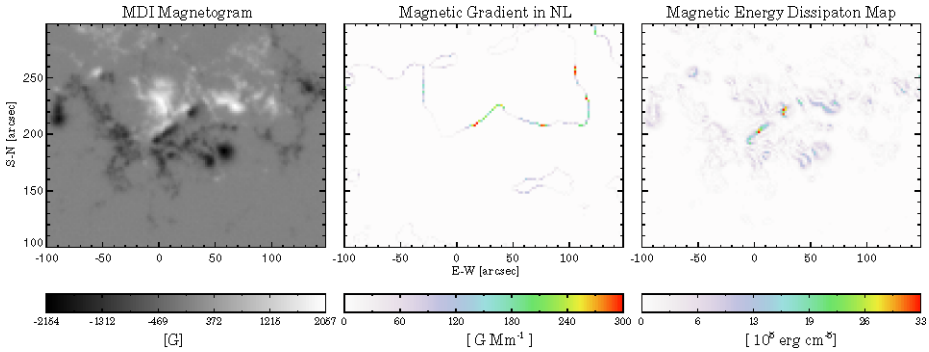

Figure 1 Left: Line-of-sight magnetogram of NOAA AR 9077 taken on 14 July 2000. Middle: Gradient distribution along the neutral line. Right: Map of the energy dissipation (Jing *et al.*, 2006). The magnitude of parameters in each pixel is indicated by the corresponding color scale bar.

of one X1.0 flare per day. The middle and right panels show the gradient distribution along the magnetic neutral line and structures of magnetic energy dissipation, respectively. The values in each pixel are indicated by the corresponding color scale bar. The quantity $L_{\rm gnl}$ is the total length of the strong gradient segments (>50 G Mm^{-1}) of the neutral line.

The majority of selected ARs produced a couple of flares with different intensities in the next 24 hours. Based on the maximum magnitude of flares they produced, ARs were classified into four levels with ordinal value 3, 2, 1, and 0. These are shown in Table 1. The first three columns show the date, the AR number, and the flare location. The next four columns show the magnetic parameters $L_{\rm gnl}$, $T_{\rm flux}$, $E_{\rm diss}$, and $F_{\rm idx}$, computed based on the previous equations. The last column named Level is our response variable to indicate the maximum magnitude of flares occurring in the following 1-day period.

2.3. Flare Statistical Characteristics

From 1998 to 2005, a total of 230 flare events are analyzed. The descriptive data for the magnetic parameters $L_{\rm gnl}$, $T_{\rm flux}$, and $E_{\rm diss}$ are summarized in Table 1. All the parameters are normalized with respect to their maxima over the entire data set so that each predictor varies between 0 and 1. In the flare event list, 34 of them (Level $= 3$) produced X-class flares, 68 (Level $= 2$) produced M-class flares, and 65 (Level $= 1$) produced C-class flares. Only a small fraction of C-class events were randomly selected to match the sample size of larger flares. For the rest of the sample (Level $= 0$), they either did not produce any flares or produced smaller flares under C-class in the period. The mean and standard deviations of each parameter are calculated and displayed according to each Level.

The mean value of $L_{\rm gnl}$ for events associated with X-class flares was found to be 118.74 Mm, much larger than that associated with either M- (64.28 Mm) or C-class (62.12 Mm) flares and an order of magnitude larger than the mean value found for those flare-quiet regions (see Table 2). Similar trends are also present for $E_{\rm diss}$ and $T_{\rm flux}$. This further evidences that the extreme events such as X-class flares have higher tendency to occur in ARs with higher concentration of free magnetic energy or nonpotentiality. For $T_{\rm flux}$, the differences among the mean values of X-, M-, and C-classes are only about 15%, not as large as for $L_{\rm gnl}$ and $E_{\rm diss}$. Flare productivity is weakly related to the active region size.

Table 1 List of active regions associated with flares.

Date	AR#	Location	L_{gnl} (403.0 Mm)	T_{flux} (1.44×10^{23} Mx)	E_{diss} (3.78×10^{9} erg cm^{-3})	F_{idx} (10^{-6} W m^{-2})	Level
17/01/2005	10721	S04E03	0.0000	0.0083	0.0073	0.01	0
23/01/2005	10726	N01W00	0.0000	0.0731	0.0588	0.01	0
02/02/2005	10729	S10W09	0.0000	0.0400	0.0366	1.63	0
08/02/2005	10731	S02W01	0.0000	0.0084	0.0086	0.01	0
02/03/2005	10739	S03W03	0.0000	0.0194	0.0162	0.84	0
15/03/2005	10743	S08W03	0.0124	0.1351	0.0858	14.58	0
02/04/2005	10747	S06W04	0.0000	0.0463	0.0559	6.84	0
08/04/2005	10749	S05E11	0.0000	0.0701	0.0582	0.01	0
11/04/2005	10750	S07E08	0.0099	0.1186	0.0981	1.25	0
08/05/2005	10758	S07E08	0.0496	0.2283	0.2223	140.88	0
04/06/2005	10769	S06E01	0.0074	0.1730	0.1312	0.80	0
10/06/2005	10775	N10E06	0.1191	0.1760	0.1482	65.48	0
04/08/2005	10796	S07W01	0.0000	0.0289	0.0196	0.17	0
18/08/2005	10798	S09E08	0.0000	0.0282	0.0223	120.27	0
07/10/2005	10813	S08E01	0.0199	0.0835	0.0840	1.40	0
20/10/2005	10815	N08E07	0.0000	0.0435	0.0257	0.01	0
02/11/2005	10819	S09W05	0.0000	0.0474	0.0290	1.71	0
03/11/2005	10818	S08W04	0.0000	0.0327	0.0183	0.20	0
26/11/2005	10825	S06E01	0.0000	0.0142	0.0085	0.01	0
15/12/2005	10834	S07W01	0.0000	0.1122	0.0755	2.76	0
29/12/2005	10840	S03E02	0.0000	0.0884	0.0594	0.01	0
13/01/1998	8131	S24W12	0.0769	0.1266	0.1214	36.44	0
11/08/1999	8662	S16E08	0.0968	0.2405	0.2228	35.64	0
19/02/2001	9354	S09W07	0.0422	0.1282	0.0933	12.49	0
10/07/2001	9531	S06E05	0.0471	0.1121	0.0821	13.40	0
18/07/2001	9545	N09E03	0.0223	0.0478	0.0279	11.26	0
20/07/2001	9542	N08E07	0.0000	0.0475	0.0230	0.49	0
31/07/2001	9557	S21E25	0.0124	0.1373	0.0890	135.02	0
08/05/2002	9937	S09E13	0.0471	0.1247	0.1082	37.63	0
13/06/2002	9991	S20E05	0.0074	0.1465	0.0993	5.44	0
18/06/2002	10000	N18E15	0.0124	0.1738	0.1351	341.48	0
04/12/2002	10208	N09E03	0.0819	0.1946	0.1800	27.48	0
05/03/2003	10296	N12E05	0.0372	0.3851	0.2814	13.12	0
12/03/2003	10306	N05E06	0.0273	0.2603	0.1603	8.25	0
15/04/2003	10334	S08E12	0.0000	0.0913	0.0685	0.55	0
17/05/2003	10357	S17E07	0.0174	0.0488	0.0466	0.85	0
25/05/2003	10365	S09E21	0.0223	0.0944	0.0984	599.27	0
20/06/2003	10388	S03E04	0.0000	0.0750	0.0818	10.67	0
09/09/2003	10456	S09E10	0.0074	0.0529	0.0531	15.26	0
06/10/2003	10471	S08E07	0.0819	0.3249	0.2594	48.83	0
12/01/2004	10537	N04W04	0.0571	0.1377	0.1168	271.55	0
24/02/2004	10564	N14E00	0.1042	0.2240	0.2290	238.04	0

(*Continued on next page*)

Table 1 (*Continued from previous page*)

Date	AR#	Location	L_{gnl} (403.0 Mm)	T_{flux} (1.44×10^{23} Mx)	E_{diss} (3.78×10^{9} erg cm^{-3})	F_{idx} (10^{-6} W m^{-2})	Level
18/05/2004	10617	S12E08	0.0000	0.0463	0.0497	3.24	0
25/05/2004	10618	S10E12	0.0868	0.3088	0.2463	107.47	0
03/06/2004	10621	S14E13	0.0422	0.1887	0.1292	6.35	0
06/06/2004	10624	S08E10	0.0000	0.0580	0.0364	0.34	0
04/08/2004	10655	S09E14	0.0769	0.1621	0.1467	10.98	0
02/10/2004	10675	S10W06	0.0000	0.0936	0.0691	0.65	0
23/10/2004	10684	S03W00	0.0223	0.1015	0.1165	0.61	0
25/11/2004	10704	N13W18	0.0223	0.1665	0.1266	2.12	0
01/12/2004	10706	S08W16	0.0000	0.1223	0.1067	27.57	0
15/02/2005	10735	S08E07	0.0521	0.2187	0.1534	13.93	0
12/03/2005	10742	S05E03	0.0571	0.2221	0.1992	25.79	0
18/04/2005	10754	S08E06	0.0124	0.0364	0.0366	0.40	0
07/05/2005	10758	S09E26	0.1241	0.2345	0.2066	140.88	0
11/06/2005	10776	S06E04	0.0620	0.2106	0.1730	37.30	0
26/07/2005	10791	N14E23	0.0571	0.0901	0.0918	7.81	0
15/08/2005	10797	S13E12	0.0298	0.1160	0.1016	0.89	0
02/11/2005	10818	S08E09	0.0000	0.0466	0.0323	0.20	0
26/11/2005	10824	S14W09	0.0000	0.0864	0.0570	10.09	0
04/12/2005	10828	S04E04	0.0074	0.0714	0.0559	1.34	0
15/12/2005	10835	N19W03	0.0099	0.0846	0.0556	7.13	0
19/12/2005	10837	S10W10	0.0124	0.0898	0.0762	2.04	0
04/11/1998	8375	N19W08	0.1687	0.2831	0.1814	220.89	1
02/06/1999	8562	S16E07	0.1663	0.1485	0.1562	21.70	1
26/06/1999	8598	N23E09	0.1390	0.5181	0.4271	71.20	1
29/06/1999	8603	S15E16	0.0099	0.3053	0.2228	77.20	1
01/07/1999	8611	S25E18	0.1464	0.2096	0.1995	160.70	1
02/08/1999	8651	N24E08	0.2010	0.5600	0.4127	153.10	1
03/08/1999	8651	N25W04	0.1638	0.5391	0.4024	153.10	1
26/08/1999	8674	S22E09	0.3474	0.8689	0.6256	346.70	1
11/11/1999	8759	N09E14	0.3251	0.5733	0.4180	113.50	1
25/11/1999	8778	S15E06	0.0695	0.1786	0.1659	138.90	1
27/11/1999	8778	S14W17	0.1117	0.2268	0.1908	138.90	1
16/03/2000	8910	N11E18	0.0868	0.2468	0.2022	437.51	1
10/04/2000	8948	S15E03	0.2333	0.3326	0.2544	216.10	1
18/04/2000	8963	N16E18	0.0298	0.1966	0.1272	54.70	1
19/04/2000	8963	N14E09	0.0720	0.1998	0.1573	54.70	1
17/05/2000	8996	S20E16	0.2258	0.5371	0.4195	129.40	1
08/06/2000	9026	N20W06	0.1613	0.4098	0.4023	945.23	1
07/07/2000	9070	N20E14	0.1886	0.2749	0.3008	186.80	1
08/07/2000	9070	N17W01	0.1836	0.2926	0.3151	186.80	1
05/09/2000	9154	S20E06	0.1538	0.2058	0.2225	55.56	1
30/09/2000	9173	S12E13	0.0819	0.2051	0.1689	50.30	1

(*Continued on next page*)

Table 1 (*Continued from previous page*)

Date	AR#	Location	L_{gnl} (403.0 Mm)	T_{flux} (1.44×10^{23} Mx)	E_{diss} (3.78×10^{9} erg cm^{-3})	F_{idx} (10^{-6} W m^{-2})	Level
09/10/2000	9182	N02W04	0.0099	0.1200	0.0582	69.50	1
31/10/2000	9209	S23W06	0.0893	0.2161	0.1655	81.20	1
22/11/2000	9236	N20E12	0.0993	0.2635	0.1476	1326.30	1
06/03/2001	9368	N26W08	0.0943	0.2840	0.2100	167.00	1
27/03/2001	9393	N18E08	0.4739	0.8200	0.7258	2954.50	1
21/05/2001	9461	N22E08	0.0273	0.2481	0.1674	18.36	1
15/07/2001	9539	S17W01	0.0893	0.1074	0.0994	60.60	1
10/09/2001	9608	S23E14	0.5732	0.7707	0.4986	498.24	1
11/09/2001	9608	S29E10	0.4268	0.6911	0.4343	498.24	1
13/09/2001	9610	S13W08	0.2010	0.4257	0.2551	31.60	1
24/09/2001	9628	S18E07	0.3027	0.6807	0.3903	274.00	1
30/09/2001	9636	N12W05	0.1663	0.4392	0.3169	100.30	1
24/10/2001	9672	S17E00	0.1315	0.3531	0.2234	475.10	1
27/10/2001	9678	N07E05	0.1241	0.2637	0.1853	103.10	1
30/10/2001	9682	N12E02	0.2159	0.4462	0.3167	269.70	1
03/11/2001	9684	N05W17	0.1290	0.2757	0.1775	145.00	1
20/11/2001	9704	S17W09	0.1663	0.4064	0.2307	283.60	1
06/01/2002	9767	S21W14	0.0323	0.5565	0.2709	61.50	1
08/01/2002	9773	N14E05	0.0943	0.3114	0.2196	290.56	1
10/01/2002	9773	N14W17	0.1811	0.4666	0.3365	290.56	1
16/07/2002	10030	N21E01	0.2233	0.5169	0.4707	793.73	1
27/07/2002	10039	S17E17	0.5136	0.8234	0.7370	733.80	1
29/07/2002	10050	S07E06	0.0695	0.2083	0.2113	60.20	1
02/08/2002	10057	S09E05	0.0496	0.0629	0.0714	72.70	1
05/09/2002	10096	N08W01	0.0596	0.3391	0.2657	23.80	1
02/10/2002	10137	S20E18	0.1191	0.1633	0.1728	174.64	1
05/11/2002	10177	N16W09	0.1340	0.2487	0.2252	80.30	1
06/11/2002	10180	S09W07	0.2407	0.3387	0.3391	259.50	1
22/02/2003	10290	N17W06	0.0769	0.1757	0.1458	36.06	1
15/03/2003	10314	S15W13	0.0620	0.1368	0.1919	529.20	1
01/05/2003	10349	S13E07	0.0893	0.3521	0.2549	86.37	1
07/06/2003	10375	N11E09	0.1514	0.3160	0.3202	1358.62	1
08/06/2003	10375	N11W03	0.1836	0.3562	0.3566	1358.62	1
18/07/2003	10410	S12E09	0.0447	0.1763	0.1692	91.71	1
15/08/2003	10431	S13W02	0.2035	0.3798	0.3812	124.65	1
28/10/2003	10488	N09W05	0.2357	0.3883	0.5545	881.80	1
25/02/2004	10564	N14W13	0.1092	0.2944	0.2537	238.04	1
29/03/2004	10582	N13E18	0.2010	0.2764	0.2162	144.65	1
31/03/2004	10582	N13W14	0.0844	0.2313	0.1605	144.65	1
19/07/2004	10649	S09W00	0.1737	0.4292	0.3259	1381.59	1
11/08/2004	10656	S14E13	0.1191	0.3580	0.3154	1260.24	1
04/06/2005	10772	S18E09	0.0397	0.0898	0.1061	98.41	1

(*Continued on next page*)

Table 1 (*Continued from previous page*)

Date	AR#	Location	L_{gnl} (403.0 Mm)	T_{flux} (1.44×10^{23} Mx)	E_{diss} (3.78×10^{9} erg cm^{-3})	F_{idx} (10^{-6} W m^{-2})	Level
02/07/2005	10785	S17E04	0.0199	0.0451	0.0431	15.56	1
14/09/2005	10808	S11E02	0.2705	0.4970	0.5185	4886.56	1
15/03/1998	8179	S24W04	0.1514	0.3054	0.2312	100.32	2
26/03/1998	8185	S24E04	0.1191	0.3059	0.1923	48.46	2
01/05/1998	8210	S17E05	0.0769	0.1908	0.0995	422.59	2
30/06/1999	8603	S14W01	0.0893	0.2998	0.2428	77.20	2
02/07/1999	8611	S26E08	0.2531	0.2536	0.2508	160.70	2
24/07/1999	8636	N20W06	0.1638	0.4340	0.3321	94.99	2
19/08/1999	8672	N16W02	0.0596	0.1164	0.0970	10.00	2
27/08/1999	8674	S21W04	0.4144	0.9085	0.6989	346.70	2
12/11/1999	8759	N10E05	0.2233	0.5727	0.4486	113.50	2
26/11/1999	8778	S14W06	0.0868	0.2062	0.1671	138.90	2
22/12/1999	8806	N19E09	0.1390	0.4728	0.3825	259.78	2
18/01/2000	8831	S17E00	0.0372	0.2686	0.1780	49.00	2
17/02/2000	8872	S28E05	0.0124	0.1196	0.0714	13.80	2
13/03/2000	8906	S17E02	0.2506	0.5409	0.3664	284.10	2
20/07/2000	9087	S12W02	0.1737	0.4648	0.3947	443.60	2
25/07/2000	9097	N06W02	0.0596	0.2461	0.1627	149.80	2
16/09/2000	9165	N15E00	0.0943	0.1973	0.1628	259.60	2
09/11/2000	9221	S12E08	0.0000	0.1329	0.0839	10.00	2
18/11/2000	9231	S21E00	0.1365	0.3023	0.2577	99.01	2
23/11/2000	9236	N22E04	0.1340	0.2945	0.2078	1326.30	2
10/01/2001	9302	N19W00	0.0372	0.3630	0.2223	56.10	2
28/03/2001	9393	N17W04	0.4739	0.8608	0.8132	2954.50	2
09/04/2001	9415	S21E04	0.1737	0.3813	0.3771	2811.82	2
25/04/2001	9433	N19E04	0.2333	0.7455	0.6640	541.09	2
05/05/2001	9445	N25W02	0.1514	0.3671	0.2907	70.80	2
13/05/2001	9455	S17E01	0.1538	0.1218	0.1766	161.04	2
04/06/2001	9484	S06E05	0.0496	0.1049	0.0801	37.00	2
03/09/2001	9601	N13E02	0.2730	0.4904	0.2930	327.31	2
25/09/2001	9628	S20E00	0.4045	0.7212	0.4767	274.00	2
29/09/2001	9636	N16E07	0.0397	0.3986	0.2554	100.30	2
06/11/2001	9687	S20E01	0.0968	0.3943	0.2367	333.10	2
10/11/2001	9690	S17E05	0.4169	0.6816	0.4435	518.83	2
11/11/2001	9690	S17W07	0.3623	0.6560	0.3822	518.83	2
29/11/2001	9715	N04E03	0.1886	0.3965	0.2904	262.60	2
09/01/2002	9773	N14W04	0.1712	0.3980	0.3168	290.56	2
14/03/2002	9866	S09E06	0.0769	0.3582	0.2569	163.70	2
15/03/2002	9866	S09W06	0.1017	0.3532	0.2742	163.70	2
10/04/2002	9893	N19W08	0.1117	0.2640	0.2084	248.70	2
15/04/2002	9906	S14W04	0.1439	0.3094	0.2365	215.82	2
28/07/2002	10039	S16E08	0.3375	0.5376	0.4921	733.80	2

(*Continued on next page*)

Table 1 (*Continued from previous page*)

Date	AR#	Location	L_{gnl} (403.0 Mm)	T_{flux} (1.44×10^{23} Mx)	E_{diss} (3.78×10^{9} erg cm^{-3})	F_{idx} (10^{-6} W m^{-2})	Level
28/07/2002	10044	S18E01	0.1464	0.4019	0.3570	309.70	2
15/08/2002	10066	N13E03	0.0298	0.1540	0.1453	22.40	2
17/08/2002	10069	S08E08	0.4020	0.6813	0.5715	1100.00	2
18/08/2002	10069	S08W07	0.4392	0.7171	0.5806	1100.00	2
23/08/2002	10083	S18W05	0.1390	0.1769	0.2134	135.80	2
03/10/2002	10137	S19E08	0.0993	0.1954	0.2050	174.64	2
04/10/2002	10137	S19W05	0.0645	0.2059	0.2142	174.64	2
25/10/2002	10162	N27W03	0.2010	0.4993	0.3659	246.48	2
16/12/2002	10227	N06W06	0.0298	0.0637	0.1110	28.30	2
17/12/2002	10226	S27W02	0.1911	0.2959	0.3312	231.60	2
19/12/2002	10229	N19W02	0.0695	0.3012	0.2240	42.30	2
07/01/2003	10244	S21W01	0.0000	0.1531	0.1213	40.17	2
23/01/2003	10266	N13W04	0.0372	0.0519	0.0702	65.81	2
21/04/2003	10338	N18E06	0.0273	0.0794	0.0678	399.41	2
24/10/2003	10484	N02E01	0.2605	0.4880	0.4011	696.70	2
18/11/2003	10501	N01E08	0.1464	0.2353	0.1946	404.78	2
19/11/2003	10501	N01W03	0.1092	0.2269	0.1833	404.78	2
18/01/2004	10540	S14E01	0.0943	0.2868	0.2208	179.69	2
23/07/2004	10652	N08E04	0.2308	0.5763	0.4502	670.64	2
12/08/2004	10656	S13E02	0.1762	0.4171	0.4164	1260.24	2
05/11/2004	10696	N09E06	0.1960	0.2312	0.2438	1120.55	2
06/11/2004	10696	N09W08	0.2382	0.2751	0.3237	1120.55	2
02/12/2004	10708	N09E01	0.0074	0.1085	0.0849	31.34	2
14/01/2005	10718	S07W08	0.0918	0.1860	0.1962	87.67	2
17/05/2005	10763	S17E06	0.1216	0.1161	0.1676	130.91	2
07/07/2005	10786	N11E08	0.1141	0.1878	0.2432	612.87	2
18/11/2005	10822	S08W01	0.0496	0.2446	0.1311	255.59	2
02/12/2005	10826	S04E06	0.0968	0.2215	0.2042	221.05	2
02/05/1998	8210	S17W12	0.1216	0.2231	0.1345	422.59	3
06/06/2000	9026	N21E18	0.1911	0.4713	0.4322	945.23	3
07/06/2000	9026	N20E05	0.1737	0.4356	0.4433	945.23	3
11/07/2000	9077	N17E45	0.4591	0.5440	0.4450	1256.40	3
12/07/2000	9077	N18E27	0.2605	0.4812	0.4272	1256.40	3
14/07/2000	9077	N17E02	0.2184	0.4170	0.4515	1256.40	3
24/11/2000	9236	N21W10	0.2134	0.3350	0.2221	1326.30	3
25/11/2000	9236	N21W24	0.1365	0.3658	0.2330	1326.30	3
29/03/2001	9393	N17W18	0.6179	0.8903	0.8736	2954.50	3
06/04/2001	9415	S21E42	0.2730	0.4382	0.3536	2811.82	3
10/04/2001	9415	S22W12	0.1787	0.3838	0.4026	2811.82	3
23/06/2001	9511	N10E23	0.0372	0.1030	0.0903	276.79	3
25/08/2001	9591	S18E40	0.4888	0.6020	0.3467	872.30	3
24/09/2001	9632	S18E28	0.3251	0.6723	0.3959	322.40	3

(*Continued on next page*)

Table 1 (*Continued from previous page*)

Date	AR#	Location	L_{gnl} (403.0 Mm)	T_{flux} (1.44×10^{23} Mx)	E_{diss} (3.78×10^{9} erg cm^{-3})	F_{idx} (10^{-6} W m^{-2})	Level
22/10/2001	9672	S19E23	0.2258	0.4881	0.2846	475.10	3
25/10/2001	9672	S19W16	0.2506	0.5218	0.2971	475.10	3
04/11/2001	9684	N05W29	0.0918	0.2919	0.1742	145.00	3
15/07/2002	10030	N19E11	0.3424	0.4931	0.4716	793.73	3
27/05/2003	10365	S06W08	0.1588	0.2344	0.2584	599.27	3
26/10/2003	10486	S16E41	1.0000	1.0000	0.6908	6829.50	3
28/10/2003	10486	S18E04	0.5831	0.9171	1.0000	6829.50	3
29/10/2003	10486	S17W09	0.6104	0.9386	0.8726	6829.50	3
26/02/2004	10564	N14W28	0.1638	0.3743	0.3001	238.04	3
15/07/2004	10649	S10E48	0.5161	0.7240	0.4931	1381.59	3
16/07/2004	10649	S08E38	0.3722	0.5656	0.4306	1381.59	3
17/07/2004	10649	S08E24	0.1911	0.4984	0.3974	1381.59	3
13/08/2004	10656	S13W12	0.2382	0.4884	0.4220	1260.24	3
30/10/2004	10691	N13W14	0.1042	0.1941	0.1842	454.48	3
07/11/2004	10696	N08W21	0.2184	0.2925	0.3263	1120.55	3
01/01/2005	10715	N04E22	0.0670	0.1295	0.1337	158.56	3
15/01/2005	10720	N13W03	0.2184	0.4803	0.4165	2379.42	3
17/01/2005	10720	N13W29	0.3375	0.5447	0.4639	2379.42	3
13/09/2005	10808	S11E17	0.4020	0.5028	0.4743	4886.56	3
15/09/2005	10808	S11W13	0.2308	0.4930	0.4978	4886.56	3

Table 2 Descriptive statistics of solar flare data.

	X class ($n = 34$)		M class ($n = 68$)		C class ($n = 65$)		B class or nonflare ($n = 63$)	
Label	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
L_{gnl} (Mm)	118.74	79.88	64.28	46.79	62.12	46.61	10.84	15.19
T_{flux} (10^{22} Mx)	7.02	3.15	5.03	2.72	4.95	2.86	1.72	1.19
E_{diss} (10^{8} G^2)	15.38	7.76	10.58	5.59	10.47	5.88	3.67	2.58

3. Ordinal Logistic Regression Model

3.1. Model Specification

There are a variety of statistical techniques that can be used to predict a response variable **Y** from a set of independent variables. Since the purpose of this paper is to estimate probabilities, the analytical technique should somehow provide it. In addition, if **Y** is categorical, with more than two categories, such a response variable essentially rules out the usual regression analysis, including a variety of linear models. The major problem with these techniques is that the linear function is inherently unbounded, whereas probabilities are bounded by 0 and 1. These make the generalized (compared with binary) logistic method the most obvious candidate for the regression analysis. It always returns values between 0 and 1. The choice of method depends on whether the response variable **Y** is measured on an ordinal or nominal scale. However, a nominal scale has categories that are not ordered. No quantitative

information is conveyed and therefore the results are qualitative rather than quantitative. Religious preference, race, and sex are all examples of nominal scales. An ordinal scale has categories that are ordered in the sense that higher numbers represent higher values. But the intervals between the numbers are not necessarily equal. Hence, an ordinal regression model is more suitable in our study since $\mathbf{Y}$ here indicates the maximum magnitude of flares the given active region may produce.

Suppose $\mathbf{Y}$ is the categorical response variable with $k+1$ ordered categories. For example,

$$\mathbf{Y} = \begin{cases} 0 = \text{weak}, \\ 1 = \text{moderately strong}, \\ \vdots = \vdots \\ k = \text{extremely strong}. \end{cases} \tag{5}$$

Let $\mathbf{X}$ denote the vector of predictive variables $\{x_1, x_2, \ldots, x_n\}$, and let $\pi_j(\mathbf{x}) = \text{Prob}(Y = j \mid \mathbf{x} = \mathbf{x})$ be the probability for the realization of $Y = j$, given $\mathbf{X} = \mathbf{x}$, $j = 0, 1, \ldots, k$. The cumulative probabilities are

$$\begin{aligned} \gamma_j(\mathbf{x}) &= \text{Prob}(Y \geq j \mid \mathbf{x} = \mathbf{x}) \\ &= \pi_j(\mathbf{x}) + \cdots + \pi_k(\mathbf{x}) \\ &= 1/\big[1 + \exp\big(-(\alpha_j + x\beta)\big)\big], \quad j = 1, \ldots, k, \end{aligned} \tag{6}$$

where $x\beta$ stands for $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$. There are k intercepts (or α's). The regression parameters α and β are estimated by the method of maximum likelihood (Agresti, 1996), which works by finding the value of β that returns the maximum value of the log-likelihood function. The expression

$$\text{Prob} = \big[1 + \exp(-x)\big]^{-1} \tag{7}$$

is called the logistic function (logit). We can solve this equation for $\alpha_j + x\beta$:

$$\alpha_j + x\beta = \ln\left[\frac{\text{Prob}}{1 - \text{Prob}}\right] = \ln[\text{odds that } Y \geq j \text{ occurs}] = \text{logit}\{Y \geq j\}. \tag{8}$$

Thus the model becomes a linear regression model in the log odds that $Y \geq j$. This is the well-known proportional odds (PO) model (McCullagh, 1980), also called the ordinal logistic model (Scott, Goldberg, and Mayo, 1997).

The logistic model formulated here for the solar flare study contains a four-state response variable. Level = 0 means the active region only produces microflares (lower than C-class flares) in the next 1-day period. Level = 1 means the active region at most produces C-class flares. Level = 2 is for M-class flares and Level = 3 is for X-class flares. Therefore, the category number $k = 3$ and predictive variables are the some for all of three magnetic parameters discussed earlier.

The model is computed with the statistical R software package (version 2.3.0 Linux system), using a procedure that supports the ordinal logistic regression model (*lrm*).[2] Ordinal logistic regression is not part of the standard R, but it can be calculated via the library *Design*[3] by using the function *lrm* (Alzola and Harrell, 2004).

[2]For details on the estimation procedure and the statistics in logistic regression models, see http://www.r-project.org/.

[3]http://biostat.mc.vanderbilt.edu/twiki/bin/view/main/design.

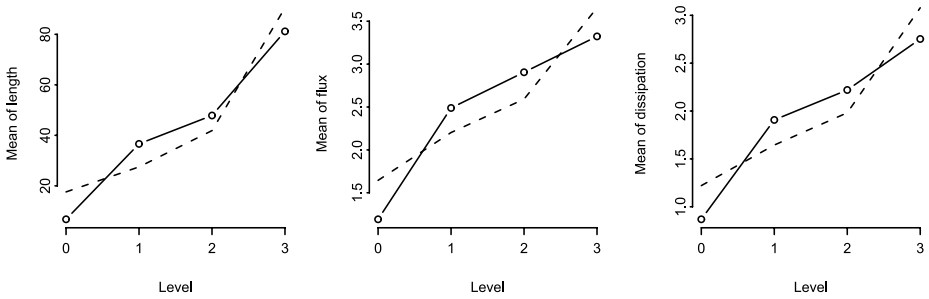

Figure 2 Examination of the ordinality of Level for every magnetic parameter by accessing how the Level is related to the mean value of each predictor, and whether the trend in each plot is monotonic. Solid lines connect the simple stratified means, and dashed lines connect the estimated expected value of $X \mid Y = j$ given that PO holds. The extent of closeness of two curves indicates the perfect condition to hold ordinal condition.

Table 3 Regression models for different combination of predictive parameters.

	Parameters	Formula
Group (a)	(1) L_{gnl}	Level $\sim L_{\text{gnl}}$
	(2) T_{flux}	Level $\sim T_{\text{flux}}$
	(3) E_{diss}	Level $\sim E_{\text{diss}}$
Group (b)	(4) $L_{\text{gnl}}, T_{\text{flux}}$	Level $\sim L_{\text{gnl}} + T_{\text{flux}} + L_{\text{gnl}} T_{\text{flux}}$
	(5) $T_{\text{flux}}, E_{\text{diss}}$	Level $\sim T_{\text{flux}} + E_{\text{diss}} + T_{\text{flux}} E_{\text{diss}}$
	(6) $L_{\text{gnl}}, E_{\text{diss}}$	Level $\sim L_{\text{gnl}} + E_{\text{diss}} + L_{\text{gnl}} E_{\text{diss}}$
Group (c)	(7) $L_{\text{gnl}}, T_{\text{flux}}, E_{\text{diss}}$	Level $\sim L_{\text{gnl}} + E_{\text{diss}} + T_{\text{flux}}$
	(8) $L_{\text{gnl}}, T_{\text{flux}}, E_{\text{diss}}$	Level $\sim L_{\text{gnl}} + T_{\text{flux}} + E_{\text{diss}} + L_{\text{gnl}} T_{\text{flux}} + T_{\text{flux}} E_{\text{diss}} + L_{\text{gnl}} * E_{\text{diss}} + L_{\text{gnl}} T_{\text{flux}} E_{\text{diss}}$

3.2. Testing for Ordinality Assumption

A basic assumption of ordinal regression models is that the response variable behaves in an ordinal fashion with respect to each predictive variable. By assuming that a predictor x is linearly related to the log odds of some appropriate event, a simple way to check for ordinality is to plot the mean of x stratified by levels of y. These means should be in a consistent order. If for many of the x's, two adjacent categories of Y do not distinguish the means, that is the evidence that the two categories of Y should be merged together.

Figure 2 displays means of all three predictive variables as calculated for each ordinal class of the response and plotted (solid) against it. In the ideal case, the dashed line (PO model) should be superposed on the solid line if the PO assumptions hold. Ordinality is satisfactorily verified for all three predictive variables (*i.e.*, they have the same monotonic trends).

3.3. Estimation Procedures

Before presenting the obtained results, we first describe three groups of models that were used in our analysis. Table 3 shows products of different data-generating models used in the

regression. To investigate the effects of each predictive parameter, every possible combination is considered.

The models in group (a) contain only one predictive parameter. For prediction purpose, these preliminary models may be too simple. However, their fitted results will help us to understand which parameter may be more significant in producing solar flares (Figure 2). Models in group (b) have three terms. The first two terms in each model are from our predictive parameters. The third one is an interaction term that exists when one independent variable change influences another one. It is also said that Variable 2 "moderates" the effect of Variable 1. In regression analysis, interaction term is quantitatively represented by the product of Variable 1 and 2. Theoretically, interactions among more than two variables, especially when these variables are continuous, can be exceedingly complex. This is because there are many different combinations of two-way interactions and because of the possibility that the order of interaction effects may be higher than two (*e.g.*, product of the square of one predictor and other predictor). Therefore, a good approach is to test for all such prespecified interaction effects with a single global test. Then, unless interactions involving only one of the predictive variables is of special interest, we can either drop all interactions or retain all of them (Harrell, 2001). The models in group (c) include all three predictive parameters, with and without corresponding interaction effect terms.

The assumption of linearity in the logistic model needs to be verified, especially when the continuous predictive variables are presented. Often, however, one of the properties of the response variable, the probability in our study, does not behave linearly in all the predictors. To test linearity, or to describe nonlinear relationships, a general way is to continuously expand predictive variables with spline functions, which are piecewise polynomials used in curve fitting. In our study, we used restricted cubic spline function (also called natural splines) with four knots on every predictive variable (Stone and Koo, 1985). For many data sets, four knots ($k = 4$) offers an adequate fit of the model and is a good compromise between flexibility and loss of precision caused by overfitting a small sample (Harrell, 2001). The locations of knots (quantiles) are fixed, when k is fixed. When $k = 4$, the quantiles are 0.5, 0.35, 0.65, and 0.95.

4. Results

4.1. Quantifying Predictive Ability of Fitted Models

A commonly used measure of predictive ability for logistic models is the fraction of correctly classified responses. One chooses a cutoff on the predicted probability of a positive response and then predicts that a response will be positive if the predicted probability exceeds this cutoff. The drawback of this method is that it is highly dependent on the cut point chosen for a positive prediction. In addition, it is a presumption to make one classification rule from a probability model.

The test statistics allow us to test whether a predictive variable, or set of variables, is related to the response. The generalized index R_N^2 (Nagelkerke, 1991; Cragg and Uhler, 1970) can be useful for quantifying the predictive strength of a model. It defined as[4]

$$R_N^2 = \frac{1 - \exp(-\mathrm{LR}/n)}{1 - \exp(-L^0/n)}, \tag{9}$$

[4]All the denotations follow the book by Harrell (2001).

Table 4 Indices to evaluate the predictive ability of models.

Models	R_N^2	c	D_{xy}
(1)	0.436	0.785	0.570
(2)	0.373	0.760	0.520
(3)	0.399	0.767	0.534
(4)	0.478	0.792	0.584
(5)	0.427	0.779	0.558
(6)	0.460	0.783	0.565
(7)	0.474	0.792	0.584
(8)	0.466	0.788	0.576

where $L^0 = -2 \times \log(\mathrm{LL})$ with LL the *likelihood*, obtained under the null hypothesis that all regression coefficients except for intercepts are zero, and n is the size of data set. The likelihood ratio (LR) is then defined as $\mathrm{LR} = L^0 - L$, where L is $-2 \times \log(\mathrm{LL_{fit}})$, achieved from the fitted model. For a perfect prediction, $2 \times \log(\mathrm{LL})$ reaches the lowest value, whereas it reaches the highest value, L^0, when a model has no predictive information. For a large enough data set,

$$R_N^2 = \frac{1-(1-\mathrm{LR}/n)}{1-(1-L^0/n)} = \frac{\mathrm{LR}}{L^0} = 1 - \frac{L}{L^0}. \tag{10}$$

Because $L < L^0$, index R_N^2 ranges from 0 to 1 and can be used to assess how well the model compares to a "perfect" model. The higher R_N^2, the more perfect the model. For the detailed mathematics, see Harrell (2001).

When cutoff points are used to define positive and negative test results in the prediction, both sensitivity and specificity are the basic measures of accuracy of the test. As the cutoff point shifts, sensitivity and specificity shift. The "receiver operating characteristic" (ROC) curve of a test is a plot of specificity (the false alarm rate) versus its sensitivity (the true positive rate) for all possible cutoff points. The area under the ROC, which is indicated by a dimensionless index c, is a convenient way of comparing predictive abilities. A c value of 0.5 indicates random predictions, whereas $c = 1$ indicates perfect prediction. In our work, models (4) and (7) have the highest value of c, 0.792 (see Table 4).

Another widely used index is Somers' index, D_{xy}, which ranks the correlation between predicted probabilities and observed responses by the difference between concordance and discordance probabilities:

$$D_{xy} = 2(c - 0.5). \tag{11}$$

When $D_{xy} = 0$, the model is making random predictions. When $D_{xy} = 1$, the predictions are perfectly discriminating.

Table 4 displays these indices for every model listed in Table 3. The models with only one predictive variable have comparable reliability in flare prediction (nearly the same indices). The indices of model (1) are slightly larger than those for models (2) and (3). The larger indices imply that the length of the strong-gradient neutral line is relatively more significant in prediction than the other two parameters. When we add one more parameter to each model, then models (4) and (6) have larger indices, indicating that the new parameter may improve the predictive strength. The worse result is for model (5) and it confirms that L_{gnl} plays the key role among the three predictors. Comparison of the results for models (7) and

(8) shows that ignoring the interaction effects between predictors improves the predictive ability. Moreover, in the overall comparison of models, models (4) and (7) stand out. Models (4) and (7), namely the combinations of $L_{\rm gnl}$, $T_{\rm flux}$, or $E_{\rm diss}$ as predictors, seem to be the most effective tools for predictions. This conclusion is consistent with the result that major flares of M or X classes are associated with a pronounced high-gradient magnetic neutral line, unsigned magnetic flux near the neutral line (Schrijver, 2007), and total magnetic energy dissipation of an active region (Jing *et al.*, 2006). Although the difference of indices of R_N^2, c, and D_{xy} from model to model is slight, the logistic ordinal regression method can still distinguish them somehow for modeling selection. This method is promising as a technique, if we come up with a better set of prediction variables to increase the best predictive power. In the current analysis, models (4) and (7) are the best candidates for predicting solar flares.

4.2. Validating the Fitted Models

Model validation is done to ascertain whether predicted values from the model are likely to accurately predict future events. The simplest validation method is one-time data-splitting. A data set is split into training (model development) and test (validation) samples by a random process. The model's calibration is validated in the test data set. One disadvantage of data-splitting is that it greatly reduces the sample size for both model development and model testing. The situation will become even worse when the original data set is not large enough, such as in our case for X-class flares. Bootstrapping can be used to obtain nearly unbiased estimates of model performance without sacrificing sample size (Efron, 1986; Breiman, 1992). With bootstrapping, one repeatedly fits the model in a bootstrap sample and evaluates the performance of the model with the original sample. The estimate of the likely performance of the final model using future data is estimated by the average of all of the indices computed using the original sample. In general, the major cause of unreliable models is overfitting the data. The amount of overfitting can be quantified by the index of overoptimism. With bootstrapping we do not have a separate validation sample for assessing calibration, but we can estimate the overoptimism by assuming that the final model needs no calibration, that is, it has overall intercept and slope corrections of 0 and 1, respectively. Refitting the model[5]

$$P_{\rm c} = \text{Prob}\{Y = 1 \mid X\hat{\beta}\} = \left[1 + \exp -(\gamma_0 + \gamma_1 X\hat{\beta})\right]^{-1}, \tag{12}$$

where $P_{\rm c}$ denotes the actual calibrated probability and the original predicted probability is $\hat{P} = [1 + \exp(-X\hat{\beta})]^{-1}$ in the original data set, will always result in $\gamma = (\gamma_0, \gamma_1) = (0, 1)$, since a logistic model is based on the overall assumption. Thus, the bias-corrected estimates of the true calibration can be obtained by the estimation of overoptimism in $(0, 1)$. An index of unreliability, $E_{\rm max}$, that represents the maximum error in predicted probabilities over the range $a \le \hat{P} \le b$, follows immediately from this calibration:

$$E_{\rm max}(a, b) = \max |\hat{P} - \hat{P}_{\rm c}|. \tag{13}$$

As an example, we validate model (7) shown in Table 3. The optimism-corrected calibrations are in Table 5. The apparent Somers' D_{xy} is 0.584, whereas the bias-corrected D_{xy} is 0.553. The slope shrinkage factor is 0.900, indicating that this model will validate using new data about 10% worse than using the current data set. The maximum absolute error

[5]The notation ˆ in this paper means quantitative expectation.

Table 5 Validations of models (4) and (7) with predictive variables L_{gnl}, T_{flux}, and E_{diss}.

	Model (4)			Model (7)		
	index.orig	optimism	index.corrected	index.orig	optimism	index.corrected
D_{xy}	0.584	0.020	0.564	0.584	0.030	0.554
R^2_N	0.478	0.031	0.447	0.474	0.045	0.430
Intercept	0.000	0.016	−0.016	0.000	−0.015	0.015
Slope	1.000	0.075	0.925	1.000	0.100	0.900
$E_{\max}$	0.000	0.020	0.020	0.000	0.026	0.026

Table 6 Validation results of all models.

Models	Bias-corrected D_{xy}	Bias-corrected R^2_N	Intercept	Slope	$E_{\max}$
(1)	0.567	0.427	−0.011	0.985	0.005
(2)	0.504	0.347	−0.100	0.944	0.014
(3)	0.529	0.386	−0.005	0.980	0.005
(4)	0.565	0.447	−0.016	0.925	0.020
(5)	0.536	0.392	−0.001	0.913	0.020
(6)	0.536	0.418	−0.020	0.910	0.024
(7)	0.553	0.430	0.015	0.900	0.026
(8)	0.557	0.436	−0.005	0.937	0.015

in predicted probability is estimated to be about 0.026. A slight decrease in R^2_N suggests some overfitting. Table 6 presents the validation results for all models. All estimates of the maximum calibration error, $E_{\max}$, are small and quite satisfactory. After the bias correction, model (7) still has the highest D_{xy} and R^2_N.

The estimated calibration curves for model (7) are displayed in Figure 3. They are calculated as

$$\text{Prob}\{\text{Level} \geq j\} = \frac{1}{1 + \exp[-(-0.0015 + 0.900 L_j)]},$$

where L_j is the logit of the predicted probability of Level $\geq j$. The closeness of the calibration curves to the bisector line demonstrates excellent validation on the absolute probability scale. The missing data in panels (a) and (c) cast some doubt on the validity of predictions for C- and X-class flares. The shape of the calibration curve in panel (b) (slope <1) implies that overfitting is present in the M-class predictions.

4.3. Describing the Fitted Models

Once the proper predictive variables have been modeled and all model assumptions have been met, it is time to present and interpret our fitted models. Equation (8) indicates that the logistic model becomes a linear model in log odds. The parameter β_j then denotes the change in the log odds per unit change in X_j, where X_j represents a single linear factor that does not interact with other variables, provided that all other variables are held constant. Instead of writing this relationship in terms of log odds, it can also be written in terms of the

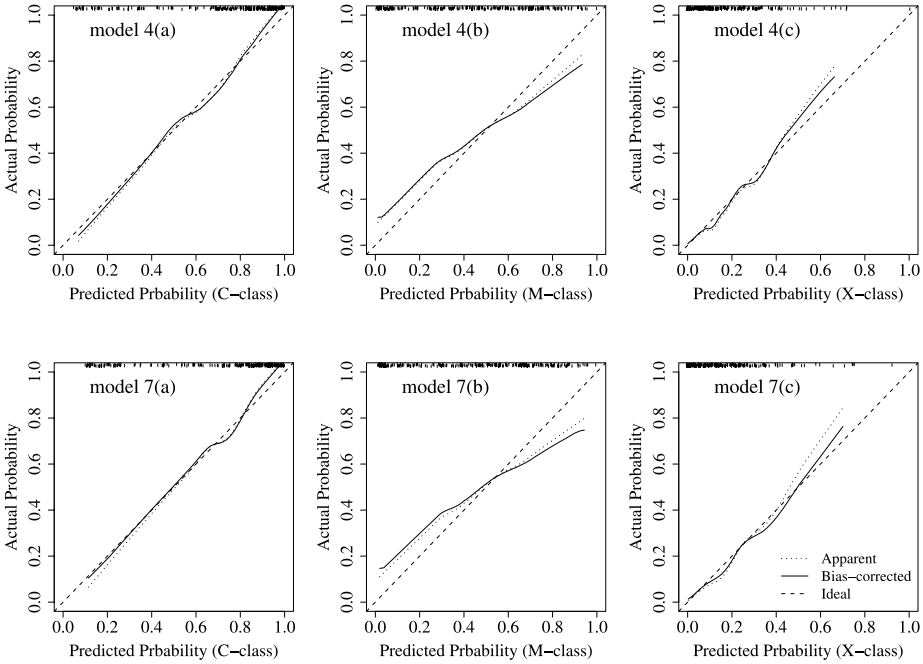

Figure 3 Estimated logistic calibration curves obtained by bootstrapping using the corrected intercept and slope. Upper panels are calculated from model (4), and lower panels from model (7). The logistic calibration model is $P_c = [1 + \exp(-(\gamma_0 + \gamma_1 L))]$, where P_c is the bias-corrected probability, L is logit($\hat{P}$), and $\hat{P}$ is the predicted probabilities (labeled as "Apparent"). The bisector line demonstrates excellent validation on an absolute probability scale.

odds that $Y \geq j$:

$$\text{odds}\{Y \geq j \mid X\} = \exp(x\beta + \alpha_j) = \exp(x\beta)\exp(\alpha_j). \tag{14}$$

The odds that $Y \geq j$, when X_j is increased by d, divided by the odds at X_j is

$$\begin{aligned} &\frac{\text{odds}\{Y \geq j \mid x_1, x_2, \ldots, x_j + d, \ldots, x_k\}}{\text{odds}\{Y \geq j \mid x_1, x_2, \ldots, x_j, \ldots, x_k\}} \\ &= \frac{\exp[\beta_j(x_j + d)]\exp(\alpha_j)}{\exp(\beta_j x_j)\exp(\alpha_j)} \\ &= \exp(\beta_j d) \end{aligned} \tag{15}$$

Thus the effect of increasing X_j by d is to increase the odds that $Y \geq j$ by a factor of $\exp(\beta_j d)$, or to increase the log odds that $Y \geq j$ by an increment of $\beta_j d$.

Table 7 contains such summary statistics for models (4) and (7). The outer quantiles of L_{gnl}, E_{diss}, and T_{flux} in model (7), again for example, are shown in the columns labeled "Low" and "High," respectively. The denotation of Δ shows the difference between "Low" and "High" quantiles. So the half-sample odds ratio for L_{gnl} is 6.84, with 0.95 confidence interval [2.23, 20.98], when T_{flux} is set to its median. The effect of increasing L_{gnl} from 0.04 (its lower quantile) to 0.19 (its upper quantile) is to increase the log odds by 1.92 or to

Table 7 Effects of L_{gnl}, E_{diss}, and T_{flux} on response variable Level for models (4) and (7).

		Low	High	Δ	Effect	Std. error	Lower 0.95	Upper 0.95
Model (4)	L_{gnl}	0.04	0.19	0.15	1.14	0.58	0.01	2.27
	Odds ratio	0.04	0.19	0.15	3.13		1.01	9.69
	T_{flux}	0.13	0.43	0.30	0.32	0.54	−0.74	1.38
	Odds ratio	0.13	0.43	0.30	1.37		0.48	3.96
Model (7)	L_{gnl}	0.04	0.19	0.15	1.92	0.57	0.80	3.04
	Odds ratio	0.04	0.19	0.15	6.84		2.23	20.98
	E_{diss}	0.12	0.34	0.22	−0.48	1.05	−2.54	1.59
	Odds ratio	0.12	0.34	0.22	0.62		0.08	4.88
	T_{flux}	0.13	0.43	0.30	1.30	0.88	−0.43	3.02
	Odds ratio	0.13	0.43	0.30	3.66		0.65	20.51

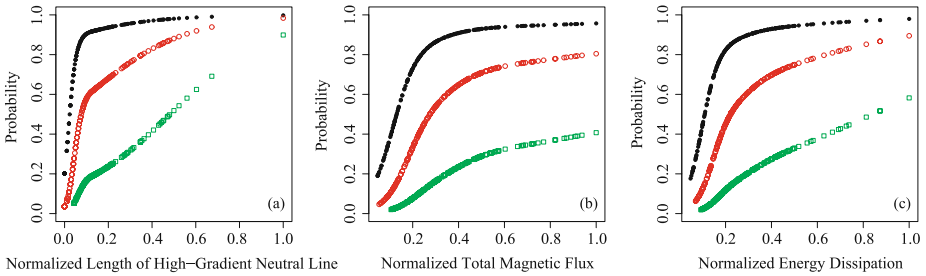

Figure 4 Distribution of predicted occurrence probability of solar flares. Panel (a), (b), and (c) show the results when only L_{gnl}, T_{flux}, or E_{diss}, respectively, is used as the predictive parameter. The probabilities for C-, M-, and X-class flares are displayed by the black dots, red circles, and green squares, respectively.

increase the odds by a factor of 6.84. The meanings of the quantities for E_{diss} and T_{flux} are the same as for L_{gnl}. Table 7 shows the relationship between an increase of the predictor's quantity and the increase of prediction probability.

Instead of displaying the result in odds, Figure 4 directly shows the predicted probabilities versus each predictive variable [models (1) – (3)]. The probability curves for C-, M-, and X-class flares are plotted in black, red, and green, respectively. The plot indicates that the occurrence probability for each class of solar flares increases with the predictive parameters and for C-class flare predictions, there is a saturation value. The C-class flare probabilities are nearly 100% when the measure values are larger than their thresholds. For M- and X-class probabilities, when L_{gnl} and E_{diss} are used as predictors [panels (a) and (c)], no such saturation value exists. The probabilities keep increasing as predictors increase. However, when T_{flux} is used to predict the probability [panel (b)], the saturation of probabilities is present for all kinds of flares. Further increase of the magnetic flux will not help to produce flares. The plot also indicates that the maximum predicted probability of X-class flares is only around 0.3 – 0.6. This may suggest that one single magnetic variable is not sufficient to predict X-class flares.

Finally, our fitted regression expression of model (7), for example, is shown as follows:

$$\text{Prob}\{\text{Level} \geq j\} = \frac{1}{1 + \exp(-\alpha_j - X\beta)}, \quad \text{where}$$

$$\hat{\alpha}_1 = -2.15,$$
$$\hat{\alpha}_2 = -4.01,$$
$$\hat{\alpha}_3 = -5.98,$$
$$\begin{aligned} X\hat{\beta} = {} & +29.55L_{\text{gnl}} - 789.30(L_{\text{gnl}})_+^3 + 1328.03(L_{\text{gnl}} - 0.06)_+^3 \\ & - 552.71(L_{\text{gnl}} - 0.16)_+^3 + 13.98(L_{\text{gnl}} - 0.43)_+^3 \\ & + 21.37E_{\text{diss}} - 365.92(E_{\text{diss}} - 0.03)_+^3 + 947.50(E_{\text{diss}} - 0.16)_+^3 \\ & - 632.49(E_{\text{diss}} - 0.27)_+^3 + 50.91(E_{\text{diss}} - 0.56)_+^3 \\ & - 9.71T_{\text{flux}} + 177.69(T_{\text{flux}} - 0.05)_+^3 - 448.94(T_{\text{flux}} - 0.19)_+^3 \\ & + 316.91(T_{\text{flux}} - 0.35)_+^3 - 45.65(T_{\text{flux}} - 0.72)_+^3, \end{aligned}$$

and $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise. L_{gnl}, T_{flux}, and E_{diss}, measured for a given active region, are then put into this equation to compute the predicted probabilities.

4.4. Comparison with NOAA/SWPC and NASA Solar Monitor Predictions

The existing methods of prediction rely on the McIntosh classification scheme of active regions (McIntosh, 1990; Bornmann and Shaw, 1994). The general expression of McIntosh classification is Zpc, where Z is the modified Zurich class, p is the type of principal spot, primarily describing the penumbra, and c is the degree of compactness in the interior of the group. According to these three components, sunspots can be classified into 60 distinct type of groups. The percentage probabilities are calculated based on the recorded number of flares produced by a given sunspot group. This approach is the basis of the prediction generated by NOAA/SWPC[6] and NASA Goddard Space Flight Center's Solar Data Analysis Center (SDAC) (Gallagher, Moon, and Wang, 2002).[7] In addition to the McIntosh classification scheme, NOAA/SWPC incorporates a lot of additional information, including dynamical properties of spot growth, magnetic topology inferred from the sunspot structure, and previous flare activity, to establish an expert system. This system involves more than 500 decision rules including those provided by human experts.

One disadvantages of the classification-based approaches is that the variation in flare probability within a class is unavoidably ignored. The classification process is somewhat subjective because the McIntosh scheme with three parameters is an arbitrary construction. Different observers may not agree with a given classification. Similar problems arise with the additional information in the expert system since the choice of properties is essentially arbitrary. Moreover, they might need human intervention, either in classification or in prediction procedures, and therefore are not suitable for automated prediction.

To compare the predictability of the logistic method [based on models (4) and (7)] and NASA/SDAC and NOAA/SWPC schemes, we studied our event list and found 55 events in the list that were also predicted by NOAA/SWPC and NASA/SDAC. Their prediction results were plotted together and shown in Figure 5. Every event (flare) is indexed on the x-axis. The Y-axis represents the predicted probability. The results from different prediction approaches are indicated by different plotting symbols. For comparison, the actual results

[6] http://www.swpc.noaa.gov/ftpdir/latest/daypre.txt.

[7] http://www.solarmonitor.org.

(1 meaning occurred, 0 meaning not) are also presented (green dots). We then used a contingency table (Table 8), which has been widely used in the meteorological forecasting literature, to evaluate the prediction capability of these approaches. This table can provide us with information on the success or failure of the forecasting experiment in real time (Kim *et al.*, 2005). We thus defined the probability of >50% to be the "yes predicted," as shown by the points above the horizontal dotted line. The vertical dotted line indicates the actual starting point of the flare. Each panel in Figure 5 is divided into four regions ($a-d$). Region a contains the events with "yes predicted" and "yes observed." Region b represents the number of false alarms ("yes predicted" but "not observed"). Similarly, c is the number of misses ("not predicted" but "yes observed") and d is the number of correct nulls ("not predicted" or "not observed").

The indices used by NOAA's National Weather Service (NWS) were computed and listed in Table 8. POD, $a/(a+c)$, is the percentage of all flare events that are predicted (with a perfect score being 100%). FAR measures how often we issue false alarm, or in other words, it is a measure of "crying wolf" $[b/(a+b)]$. Ideally we want this number to be 0.0%. CSI is the ratio of predicted events a to the total number of $(a+b+c)$. In C- and M-class flare prediction, the predicted probabilities computed from NASA/SDAC present weak capabilities. The "yes predicted" is not as obvious as those from the other two methods. Meanwhile, the minimum probabilities predicted by the logistic method are comparable with the results from NOAA/SWPC. For X-class prediction, the results from all current methods are not satisfactory. Thus, the indices show that the method used by NOAA/SWPC provides the best prediction results. The low predictability in forecasting X-class flares perhaps indicates that the predictive parameters we applied so far may not have strong enough correlation in triggering X-class flares or is due to the insufficient data samples in the logistic regression model.

The gap between NOAA/SWPC and the logistic regression model, especially model (4), is quite small when forecasting major solar flares. In Figure 5, the probabilities of X-class flare prediction obtained from the ordinal logistic method and NOAA/SWPC are lower than the cutoff of 50%. We therefore downgrade the cutoff probability to 25% and recount the values of a, b, c, and d. The obtained indices are displayed in the last three columns of Table 8. Each index of X-class flare prediction from the logistic method is almost identical with the one from NOAA/SWPC. We propose that the ordinal logistic method is a promising tool in forecasting major flares, especially as we will assemble enough data samples and even more predictive parameters in the future.

5. Conclusions

In this paper we propose a statistical ordinal logistic regression model to solar flare prediction. For this, we selected 174 active regions from 1996 to 2005, computed their corresponding magnetic parameters $L_{\rm gnl}$, $T_{\rm flux}$, and $E_{\rm diss}$ measured from SOHO MDI magnetograms, and then applied our logistic model to them. Our main results can be summarized as follows:

1. Each photospheric magnetic parameter – $L_{\rm gnl}$, $T_{\rm flux}$, and $E_{\rm diss}$ – has a positive correlation with the predicted probability. Among them the most significant variable is $L_{\rm gnl}$, followed by $T_{\rm flux}$ and $E_{\rm diss}$.
2. The ordinal logistic regression method is proven to be a viable, practical, and competitive approach to automated flare prediction. The results are better than those data published by the NASA/SDAC service and are comparable to the data provided by the NOAA/SWPC complicated expert system.

Table 8 Comparison among three prediction approaches.

	C-class flare prediction				M-class flare prediction				X-class flare prediction[1]		
	Logistic				Logistic				Logistic		
	Model (4)	Model (7)	NOAA/SWPC	NASA/SDAC	Model (4)	Model (7)	NOAA/SWPC	NASA/SDAC	Model (4)	Model (7)	NOAA/SWPC
a: yes predicted	17	18	18	14	11	11	12	2	5	5	4
b: false alarms	7	14	5	2	1	1	2	1	1	1	2
c: misses	3	2	1	6	6	6	4	14	2	3	3
d: correct nulls	28	2	31	33	37	37	37	38	47	46	46
POD: $a/(a+c)$	0.85	0.90	0.95	0.70	0.65	0.65	0.75	0.13	0.71	0.63	0.57
FAR: $b/(a+b)$	0.29	0.44	0.22	0.13	0.08	0.08	0.14	0.33	0.17	0.17	0.33
CSI: $a/(a+b+c)$	0.63	0.53	0.75	0.64	0.61	0.61	0.67	0.12	0.63	0.55	0.44

[1]In X-class flares prediction, a, b, c, d are redefined by the new cutoff probability $>25\%$.

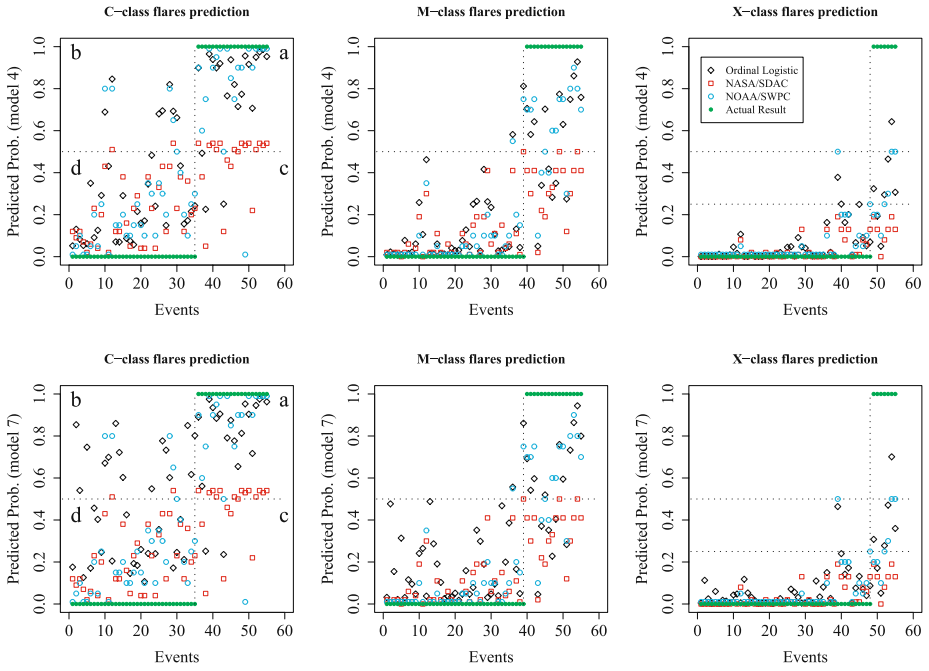

Figure 5 Comparison of three predictive methods for each level of solar flares. The upper panels are derived from model (4) and the lower panels are from model (7) (see text). The results from the ordinal logistic method, NASA/SDAC, and NOAA/SWPC are indicated by black diamonds, red squares, and blue circles, respectively. For comparison, the actual probabilities of producing flares are shown by green dots. The horizontal dotted line is the probability of 50% (25% in the X-class panel). The vertical dotted line represents the turning point of flare occurrence.

3. The prediction capabilities are different from model to model, even for the best models (4) and (7). Model (4) is a combination with $L_{\rm gnl}$, $T_{\rm flux}$ and their interaction, whereas model (7) includes all three parameters ($L_{\rm gnl}$, $T_{\rm flux}$, and $E_{\rm diss}$) without interaction effects. Our results show that model (4) might be better than model (7). Probably, it is suggested that $E_{\rm diss}$ is not that strongly related with flares. To our knowledge, this is the first time that a logistic regression model has been introduced into solar physics for flare prediction. Besides Wheatland's (2004) work, our study is another method to quantify the occurrence probability of flares into a mathematical expression.
4. According to the results from the contingency table (Table 8), all three approaches can get good results in forecasting C-class flares (with CSI between 0.53 and 0.75). In the M-class prediction, only the logistic and NOAA/SWPC approaches are feasible (0.61 and 0.67, respectively). For X-class flare prediction, the 50% cutoff is too strict for all methods to achieve. This perhaps implies that the current parameters used in prediction are insufficient to forecast X-class flares. After we changed the cutoff probability to 25%, both methods appear acceptable. The ordinal logistic method performed well in flare prediction.

So far our prediction model is limited to those magnetic parameters obtained only through SOHO MDI magnetograms. There are several physical parameters that are considered to improve the forecast capability of solar flares. These parameters need to be derived from the vector magnetograms. It has been suggested that the occurrence of flares is related to (1) the

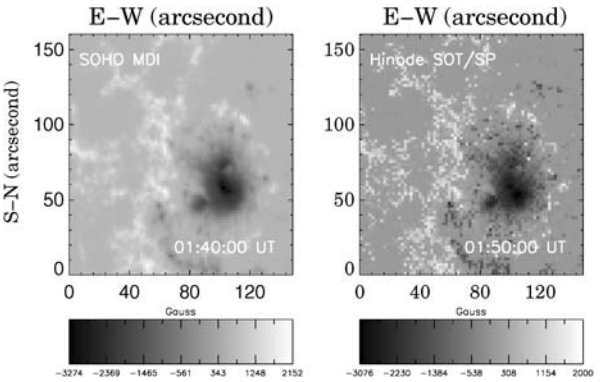

Figure 6 Comparison of observations of AR10953 on 1 May 2007 from MDI and *Hinode*. The right panel is the adjusted MDI magnetogram and the left panel is the *Hinode* SOT/SP magnetogram.

Table 9 Comparison of the measurements of AR10953 from MDI and *Hinode* $L_{\rm gnl}$, $T_{\rm flux}$, and $E_{\rm diss}$.

	$L_{\rm gnl}$ (Mm)	$T_{\rm flux}$ (10^{22} Mx)	$E_{\rm diss}$ (10^9 G^2)
Hinode (smoothed)	60.00	3.82	1.87
MDI (raw)	40.00	2.90	0.82
MDI (corrected)	55.00	4.17	1.68
Diff. (corrected MDI vs. *Hinode*)	8.3%	9.1%	10.2%

length of the strongly sheared magnetic neutral line (Falconer, Moore, and Gary, 2003), (2) the total unsigned vertical current $\int |J_z| \, {\rm d}A$, where J_z is the vertical current density, and (3) the magnetic free energy $\int \rho_{\rm e} \, {\rm d}v$, where $\rho_{\rm e}$ is the volume density of the magnetic free energy (Wang *et al.*, 1996, Leka and Barnes, 2003a, 2003b). More extensive investigation is in preparation as these parameters become readily available in the near future when the *Solar Dynamics Observatory* is launched.

Acknowledgements The authors are grateful to the referee for valuable comments and suggestions. They also thank Dr. T.E. Berger for providing the method to correct the underestimation problem of MDI. SOHO is a project of international cooperation between NASA and ESA. This work is supported by NSF under Grant Nos. IIS-0324816, ATM-0548952, ATM-0342560, and ATM-0536921 and by NASA under Grant No. NNG0-6GC81G. VY's additional support is from NASA Grant Nos. NNG0-5GN34G and NASA ACE NNG0-4GJ51G.

Appendix

To estimate the differences in measurements of magnetic flux density in space-borne data, we compared the observations from MDI and *Hinode*. Figure 6 presents the line-of-sight magnetograms of AR10953 on 1 May 2007. Berger and Lites (2003) showed that magnetic flux densities in MDI data are underestimated and the linear flux density underestimate factors derived in their analysis are roughly 0.65 in plage and 0.70 in sunspots (umbra and penumbra) or active regions. The underestimate of the MDI magnetograms is significant, so that a correction using Equation (1) must be made. Table 9 shows the comparison of the three magnetic parameters before and after correction. All the differences are smaller than or around 10%.

References

Abramenko, V.I.: 2005, *Astrophys. J.* **629**, 1141.
Abramenko, V.I., Gopasyuk, S.I., Ogir', M.B.: 1991, *Solar Phys.* **134**, 287.
Abramenko, V.I., Yurchyshyn, V.B., Wang, H., Spirock, T.J., Goode, P.R.: 2003, *Astrophys. J.* **597**, 1135.
Agresti, A.: 1996, *An Introduction to Categorical Data Analysis*, Wiley, New York
Alzola, C.F., Harrell, F.E.: 2004, *An Introduction to S and the Hmisc and Design Libraries*, Free available electronic book.
Antalova, A.: 1996, *Contrib. Astron. Obs. Skaln. Pleso* **26**, 98.
Berger, T.E., Lites, B.W.: 2003, *Solar Phys.* **213**, 213.
Bornmann, P.L., Shaw, D.: 1994, *Solar Phys.* **150**, 127.
Breiman, L.: 1992, *J. Am. Stat. Assoc.* **87**, 738.
Brueckner, G.E., Delaboudiniere, J.P., Howard, R.A., Paswaters, S.E., St. Cyr, O.C., Schwenn, R., Lamy, P., Simnett, G.M., *et al.*: 1998, *Geophys. Res. Lett.* **25**, 3019.
Cane, H.V., Richardson, I.G., St. Cyr, O.C.: 2000, *Geophys. Res. Lett.* **27**, 3591.
Canfield, R.C., Hudson, H.S., McKenzie, D.E.: 1999, *Geophys. Res. Lett.* **26**, 627.
Cragg, J.G., Uhler, R.: 1970, *Can. J. Econ.* **3**, 386.
Efron, B.: 1986, *J. Am. Stat. Assoc.* **81**, 461.
Falconer, D.A.: 2001, *J. Geophys. Res.* **106**, 25185.
Falconer, D.A., Moore, R.L., Gary, G.A.: 2003, *J. Geophys. Res.* **108**, SSH11.
Gallagher, P.T., Moon, Y.J., Wang, H.: 2002, *Solar Phys.* **209**, 171.
Gary, G.A., Hagyard, M.J.: 1990, *Solar Phys.* **126**, 21.
Gopalswamy, N., Lara, A., Lepping, R.P., Kaiser, M.L., Berdichevsky, D., St. Cyr, O.C.: 2000, *Geophys. Res. Lett.* **27**, 145.
Hagyard, M.J., Teuber, D., West, E.A., Smith, J.B.: 1984, *Solar Phys.* **91**, 115.
Harrell, F.E.: 2001, *Regression Modeling Strategies with Application to Linear Models, Logistic Regression, and Survival Analysis*, Springer, Berlin.
Hess, W.N. (ed.): 1964, *The Physics of Solar Flares*, NASA, Washington.
Jing, J., Song, H., Abramenko, V.I., Tan, C., Wang, H.: 2006, *Astrophys. J.* **644**, 1273.
Kim, R.S., Cho, K.S., Moon, Y.J., Kim, Y.H., Yi, Y., Dryer, M., Bong, S.C., Park, Y.D.: 2005, *J. Geophys. Res.* **110**, 11104.
Leka, K.D., Barnes, G.: 2003a, *Astrophys. J.* **595**, 1277.
Leka, K.D., Barnes, G.: 2003b, *Astrophys. J.* **595**, 1296.
Leka, K.D., Canfield, R.C., McClymont, A.N., de La Beaujardiere, J.F., Fan, Y., Tang, F.: 1993, *Astrophys. J.* **411**, L370.
McCullagh, P.: 1980, *J. Roy. Stat. Soc. B* **42**, 109.
McIntosh, P.S.: 1990, *Solar Phys.* **125**, 251.
Moreton, G.E., Severny, A.B.: 1968, *Solar Phys.* **3**, 282.
Nagelkerke, N.J.D.: 1991, *Biometrika* **493**, 205–247.
Priest, E., Forbes, T.: 2000, *Magnetic Reconnection: MHD Theory and Applications*, Cambridge Univ. Press, Cambridge.
Sawyer, C., Warwick, J.W., Dennett, J.T.: 1986, *Solar Flare Prediction*, Colorado Assoc. Univ. Press, Boulder.
Schrijver, C.J.: 2007, *Astrophys. J.* **662**, L119.
Scott, S.C., Goldberg, M.S., Mayo, N.E.: 1997, *J. Clin. Epidemiol.* **50**, 45.
Song, H., Yurchyshyn, V., Yang, G., Tan, C., Chen, W., Wang, H.: 2006, *Solar Phys.* **238**, 141.
Stone, C.J., Koo, C.Y.: 1985, In: *Proceeding of the Statistics Computing Section ASA, Washington, DC* 45.
Svestka, Z.: 1976, *Solar Flares*, Reidel, Dordrecht
Tan, C., Jing, J., Abramenko, V.I., Pevtsov, A.A., Song, H., Park, S.-H., Wang, H.: 2007, *Astrophys. J.* **665**, 1460.
Tian, L., Liu, Y., Wang, J.: 2002, *Solar Phys.* **209**, 361.
Wang, H.: 2006, *Astrophys. J.* **649**, 490.
Wang, T., Xu, A., Zhang, H.: 1994, *Solar Phys.* **155**, 99.
Wang, J., Shi, Z., Wang, H., Lü, Y.: 1996, *Astrophys. J.* **456**, 861.
Wang, Y.M., Ye, P.Z., Wang, S., Zhou, J.P., Wang, J.: 2002, *J. Geophys. Res.* **107**, 2.
Wang, H., Song, H., Yurchyshyn, V., Deng, Y., Zhang, H., Falconer, D., Li, J.: 2006, *Chin. J. Astron. Astrophys.* **6**, 477.
Webb, D.F., Cliver, E.W., Crooker, N.U., Cry, O.C.S., Thompson, B.J.: 2000, *J. Geophys. Res.* **105**, 7491.
Wheatland, M.S.: 2004, *Astrophys. J.* **609**, 1134.
Zhang, J., Dere, K.P., Howard, R.A., Bothmer, V.: 2003, *Astrophys. J.* **582**, 520.