Statistical Assessment of Photospheric Magnetic Features in Imminent Solar Flares Predictions

Hui Song (hxs6800@njit.edu), Vasyl Yurchyshyn , Ju Jing, Changyi Tan , V. I. Abramenko and Haimin Wang

Center for Solar-Terrestrial Research,

New Jersey Institute of Technology, Newark, NJ 07102 U.S.A

Big Bear Solar Observatory,

40386 North Shore Lane, Big Bear City, CA 92314 U.S.A.

Abstract. In this study we used 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. Three predictive parameters are: (1) total unsigned magnetic flux $T_{\rm flux}$, which is a measure of an active region's size, (2) the length of strong-gradient neutral line $L_{\rm gnl}$, which describes the global non-potentiality of an active region, and (3) total magnetic dissipation $E_{\rm diss}$, which is another proxy measure of an active region's non-potentiality. They are all derived from SOHO MDI magnetograms. The ordinal response variable is the different level of solar flares magnitude. By analyzing 230 active regions, $L_{\rm gnl}$ is proved to be the most powerful predictor, if only one predictor is chosen. Compared with the current predictions methods used by Solar Mornitor at Solar Data Analysis Center (SDAC) and NOAA Space Environment Center (SEC), the ordinal logistic model using $L_{\rm gnl}$ and $T_{\rm flux}$ as predictors demonstrated its automaticity, simpleness and fairly high prediction accuracy. To our knowledge, this is the first time the ordinal logistic regression model was used in solar physics to predict solar flares.

1. Introduction

Over the past decades, mankind has become more and more dependent on space systems, satellite-based services, as well as various ground-based facilities. All these technologies are influenced by Sun-Earth interaction phenomena. 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 *et al.*, 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 *et al.* 1986). Canfield *et al.* (1999) showed that CMEs also tend to arise in connection with active regions (ARs) exhibiting strong sheared and/or twisted coronal loops called sigmoid. 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 due 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 *et al.*, 1991; Leka *et al.*, 1993; Wang *et al.*, 1994; Wang *et al.*, 1996; Tian *et al.*, 2002). Through five solar flares, Wang (2006a) found there are obvious changes of the magnetic gradient occurred immediately and rapidly following the onset of each flare. Falconer *et al.* (2001, 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 6 large (X5 or larger) flares, Wang *et al.* (2006b) 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 strong gradient neutral line, $L_{\rm gnl}$, was proved to be a viable tool to locate source regions of either CMEs or flares. The overall accuracy of this method is about 75 % (55 out of 73 events). Jing et al. (2006) analyzed three magnetic parameters: i) mean spatial magnetic field gradient at strong-gradient magnetic neutral line, Mgnl ; ii) length of a strong-gradient magnetic neutral line, $L_{\rm gnl}$; and iii) total magnetic energy dissipated in a unit layer in 1 second over the active regions area, $E_{\rm diss}$, and found that these parameters have a positive correlation with the overall flare productivity of ARs. ARs 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, 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 $L_{\rm gnl}$, $E_{\rm diss}$, that were used in Jing *et al.* (2006) study 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 the certain response value (0, 1, 2 or 3). Thus, in this study, *Prob* represents the probability of 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 if the ordinal logistic regression model is able to predict the occurrence of solar flares in the next 1-day period. The reminder 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 US National Oceanic and Atmospheric Administration (NOAA) space environment center (SEC)[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 occurred between 1996 to 2005. The criteria for flares selection are: (1) the location of the flare is accurately indicated in the reports and as close to disk center as possible ($\pm 40^\circ$ in longitude and $\pm 30^\circ$ in latitude), so the project effects of magnetic fields can be avoided. In order to have enough events number of X-class flares for our statistical study, the longitude was extended to $\pm 40^\circ$; (2) Michelson Doppler Imager (MDI) full disk magnetograms on board Solar and Heliospheric Observatory (SOHO) was available. These magnetograms were used to analyze photospheric magnetic parameters. The reason that we use only MDI magnetograms is primarily because these data are routinely obtained, extensively achieved and free of the atmospheric seeing. Total 230 solar flare events were chosen to be analyzed.

2.2. Definition of the Predictive and Response Variables

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

1. Total unsigned magnetic flux, $T_{\rm flux}$, is a measure of the active region's size.

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

2. Length of the strong-gradient neutral line, $L_{\rm gnl}$, describes the global non-potentiality of an active region. The spatial gradient is calculated as

$$\nabla B_z = \left[\left(\frac{dB_z}{dx} \right)^2 + \left(\frac{dB_z}{dy} \right)^2 \right]^{1/2} \tag{1}$$

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 *et al.*, 2003; Song *et al.*, 2006).

3. Total magnetic energy dissipation of B_z in a unit layer per unit time, $E_{\rm diss} = \int \varepsilon(B_z) dA$, where the summation is done over the entire active region area A. The $\varepsilon(B_z)$ is defined according to the following expression (Abramenko, *et al.* 2003):

$$\varepsilon(B_z) = \left(4 \left[\left(\frac{dB_z}{dx} \right)^2 + \left(\frac{dB_z}{dy} \right)^2 \right] + 2 \left(\frac{dB_z}{dx} + \frac{dB_z}{dy} \right)^2 \right), \tag{2}$$

According to Abramenko *et al.* (2003), this measure indicates the energy dissipated at very small scales (2-3 Mm) due to the turbulent motions of magnetic flux tubes in the photosphere. Due to the gradient of B_z is also included in $\varepsilon(B_z)$, it could be another proxy measure of an active region's non-potentiality.

4. Overall flare productivity $F_{\rm idx}$ of a given active region, which is quantified by the 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).

$$F_{idx} = (100 \times \sum_{\tau} I_X + 10 \times \sum_{\tau} I_M + 1 \times \sum_{\tau} I_C + 0.1 \times \sum_{\tau} I_B)/\tau \tag{3}$$

where τ is the length of time (measured in days) during which an active region is visible on the solar disk, I_X, I_M, I_C and I_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 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 2000 July 14. The left panel shows the MDI line-of-sight magnetogram of this flare active region. The overall $F_{\rm idx}$ is as high as 1256.40 (in units of 10^{-6}Wm^{-2}), equivalent to a specific flare productivity of one super 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 couple of flares with different intensities in the next 24 hours. Based on the maximum magnitude of flares they produced, ARs were classified into 4 levels with ordinal value 3, 2, 1 and 0. They are shown in Table I. First three columns show date, the AR number, and the flare location. The next four columns show 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 occurred in the next following 1-day period.

$T_{\rm flux}$, $L_{\rm gnl}$ and $E_{\rm diss}$ parameters were proved earlier to be positively correlated with the $F_{\rm idx}$ of the next 1-day (Jing, *et al.*, 2006) and confirmed again in Figure 2. This figure has 16 scatter plots placed in 4 rows and 4 columns, each one corresponding to one of the four observed variables. Each plot in this matrix shows a scatterplot of two variables. The matrix is symmetric about its diagonal. The correlation coefficients (CCs) between $L_{\rm gnl}$, $T_{\rm flux}$, $E_{\rm diss}$ and $\log_{10}(F_{\rm idx})$ CCs varies in the range of 0.60 to 0.65. The CCs indicates that they could be used as predictive variables, either individually or combinationally, in flare forecasting.

2.3. FLARES STATISTICAL CHARACTERISTICS

From 1998 to 2005, a total of 230 flare events analyzed. The descriptive data for the magnetic parameters $L_{\rm gnl}$, $T_{\rm flux}$ and $E_{\rm diss}$ are summarized in Table II. Among 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 small fraction of C-class events were randomly selected to match the sample size of larger flares. For the left (*Level*=0), they either did not produce any flares or produced smaller flares under C-class in the next 1-day period. According to each *Level*, the mean and standard deviations of each parameter are calculated and displayed.

Mean value of $L_{\rm gnl}$ for events, associated with X-class flares, was found to be 81.18 Mm, much larger than that associated with either M- (47.86 Mm) or C-class (36.62 Mm) flares and an order of magnitude larger than the mean value found for those flare-quiet regions. The same trend is also present in the case of $E_{\rm diss}$. This further evidences that the extreme events such as X-class flares have higher tendency to occur in the ARs with high concentration of free magnetic energy. As to $T_{\rm flux}$, the differences between the mean values of X-, M- and C-class are only about 15%, not as large as for $L_{\rm gnl}$ and $E_{\rm diss}$. Flare productivity is only weakly related to the active region size.

3. Ordinal Logistic Regression Model

3.1. Model Specification

There is 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 usual regression analysis, including the variety of linear models. The major problem with these techniques is that the linear function is inherently unbounded, while probabilities are bounded by 0 and 1. This make the generalized (compared with binary) logistic method the most obvious candidates for the regression analysis. It always returns values between 0 and 1. Depending on the scale of **Y** (ordinal or nominal), the model is further classified into ordinal regression and nominal regression model. In our study, we use ordinal regression model since **Y** here indicates the maximum magnitude of flares the given active region may produce.

Suppose **Y** is the categorical response variable with *k+1* ordered categories. For example

$$\mathbf{Y} = \begin{cases} 0 = & \text{weak} \\ 1 = & \text{moderately strong} \\ \vdots = & \vdots \\ k = & \text{extremely strong} \end{cases} \tag{4}$$

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

$$\begin{aligned} \gamma_j(\mathbf{x}) &= Prob(Y \geq j|\mathbf{x} = \mathbf{x}) \\ &= \pi_j(\mathbf{x}) + \cdots + \pi_k(\mathbf{x}) \\ &= 1/[1 + exp(-(\alpha_j + x\beta))], \qquad j = 1, \cdots, k, \end{aligned} \tag{5}$$

where $x\beta$ stands for $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$. There are k intercepts (αs). 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. Expression

$$Prob = [1 + exp(-x)]^{-1} \tag{6}$$

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

$$\alpha_j + x\beta = log[\frac{Prob}{1 - Prob}] = log[\text{odds that Y} \geq \text{j occurs }] = \text{logit}\{\text{Y} \geq \text{j}\}. \tag{7}$$

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 ordinal logistic model (Scott, *et al.*, 1997).

The logistic model formulated here for the solar flares study, contains a four-state response variable. $Level = 0$ means the active region only produce microflares (Lower than C-class flares) in the next 1-day period. $Level = 1$ means the active region at most produce C-class flares. $Level = 2$ is for M- and $Level = 3$ is for X-class flares. Therefore, the category number k = 3 and predictive variables are the some or 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 ordinal logistic regression model (*lrm*). For details on the estimation procedure and the statistics in logistic regression models, see website[2]. Ordinal logistic regression is not part of the standard R, but can be calculated via library *Design*[3] by using function *lrm* (Alzola and Harrell, 2004).

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. 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 those levels of Y should be pooled.

Figure 3 is such displays. Means of all three predictive variables are calculated for each ordinal class of the response and plotted (solid) against it. In the ideal case, the dotted line (PO model) should be superposed on the solid line if the PO assumptions hold. Ordinality is satisfactorily verified for all three predictive variables (same monotonic trends).

Figure 4 shows another way to assess the PO assumption. Each predictive variable is categorized into quartiles. Each quartile group is identified using the upper and lower endpoints within that quartile. The logits of all proportions of the Level $\geq j, j = 1,2,3.$ is computed. When proportional odds holds, the difference in logits between different values of j should be the same at all levels of each parameter. This is because the model dictates that logit(Level $\geq$ j|x) - logit(Level $\geq$ i|x) = α_j - α_i, for any constant x.

3.3. Estimation Procedures

Before presenting the obtained results, we first describe three groups of models that were used in our analysis. Table III shows products of different data generating models used in the regression. In order to investigate the effects of each predictive parameter, every possible combination is analyzed.

[2] `http://www.r-project.org/`

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

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. Models in group (b) have three terms. The first two terms in each model are from our predictive parameters. The third one is called the interaction term. It exists when the effect of one independent variable changes with different values of another independent variable. 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 the possibility of 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 need to be verified, especially when the continuous predictive variables are presented. Often, however, the property of 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 expand predictive continuous 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 4 knots on every predictive variable (Stone and Koo, 1985). For many datasets, 4 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 cutpoint chosen for a positive prediction. In addition, it is presumptions 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. Let us assume that the log likelihood (L.L.) of a model is represented by:

$$R_N^2 = \frac{1 - exp(-LR/n)}{1 - exp(-L^0/n)}, \tag{8}$$

where $L^0 = -2 \times L.L$, obtained under the null hypothesis that all regression coefficients except for intercepts are zero. Likelihood ratio (LR) is then $L^0 - L$, where L is -2×L.L, achieved from the fitted model. n is the size of dataset. For large enough datasets, LR approximately follows χ^2 distribution. Index R_N^2 ranges from 0 to 1 and can be used to assess how well the model compares to a "perfect" model.

A dimensionless index c indicates probability of concordance between the predicted probability and the response. It has been shown that c is identical to a widely used measure of diagnostic discrimination, which is the area under a "receiver operating characteristics" (ROC) curve. A value of c 0.5 indicates random predictions, while c=1 indicates perfect prediction. A model that has c near 0.8 has some reliability in predicting the responses of individual events.

Another widely used index is 'Somers' index, $D_{\rm xy}$, that ranks the correlation between predicted probabilities and observed responses by the difference between concordance and discordance probabilities.

$$D_{\rm xy} = 2(c - 0.5). \tag{9}$$

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

Table IV displays these indexes for every model listed in Table III. For the models with only one predictive variable, they have comparable reliability in flare prediction (nearly same indexes). The indexes of model (1) are slightly larger than that for model (2) and (3). The larger indexes implies that the length of strong gradient neutral line is relative more significant in prediction than the other two parameters. When we add one more parameter to each model, then model (4) and (6) have larger indexes, indicating the new parameter may improve the predictive strength. The worst result is for model (5) and it confirms that $L_{\rm gnl}$ plays the key role among three predictors. The nearly same results for model (7) and (8) show that the ignorance of the interaction effects between predictors does not reduce the predictive ability. Moreover, from the comparison of models (4) and (7), it follows that parameter $E_{\rm diss}$ may be the least effective in flare prediction, while model (4), namely the combination of $L_{\rm gnl}$ and $T_{\rm flux}$ as predictors, seems to be the most effective tool for predictions. This conclusion is consistent with the result that major flares of class

M or X are associated with pronounced high-gradient magnetic neutral line (Schrijver, 2007). According to Schrijver (2007), the measure of unsigned magnetic flux near the neutral line is proved to be related with the probability of a active region to produce major flares.

4.2. VALIDATING THE FITTED MODELS

Model validation is done to ascertain whether predicted values from the model are likely to accurately predict responses on future events. The simplest validation method is one time data-splitting. A dataset is split into training (model development) and test (validation) samples by a random process. The model's calibration are validated in the test dataset. 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 dateset is not large enough, like our case in 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 on the original sample. The estimate of the likely performance of the final model on future data is estimated by the average of all of the indexes computed on 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 in 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

$$P_c = \text{Prob}\left\{\text{Y} = 1 \mid \text{X}\hat{\beta}\right\} = [1 + \exp -(\gamma_0 + \gamma_1 \text{X}\hat{\beta})]^{-1}, \tag{10}$$

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

$$E_{\max}(a, b) = max \mid \hat{P} - \hat{P}_c \mid . \tag{11}$$

As an example, we first validate model (4) shown in Table III. The optimism-corrected calibrations are in Table V. The apparent Somers' D_{xy} is 0.579, while the bias-corrected D_{xy} is 0.559. The slope shrinkage factor is 0.933, indicating that this model will validate on new data about 6.7% worse than on the current dataset. The maximum absolute error in predicted probability is estimated to be about 0.017. A slight decrease in R^2 suggests some overfitting. Table VI 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 (4) still has the highest D_{xy} and R^2.

The estimated calibration curves for model (4) are displayed in Figure 5. They are calculated as:

$$\text{Prob}\{\text{Level} \geq \text{j}\} = \frac{1}{1+\exp[-(-0.009+0.933\text{L}_\text{j})]},$$

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 panel (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 modelled and all model assumptions have been met, it is the time to present and interpret our fitted models. Equation (7) 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 odds that $\text{Y} \geq \text{j}$:

$$\text{odds}\{\text{Y} \geq \text{j} \mid \text{X}\} = \exp(\text{x}\beta + \alpha_\text{j}) = \exp(\text{x}\beta)\exp(\alpha_\text{j}). \tag{12}$$

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

$$\begin{aligned} & \frac{\text{odds}\left\{\text{Y} \geq \text{j} \mid \text{x}_1, \text{x}_2, \cdots, \text{x}_\text{j}+\text{d}, \cdots, \text{x}_\text{k}\right\}}{\text{odds}\left\{\text{Y} \geq \text{j} \mid \text{x}_1, \text{x}_2, \cdots, \text{x}_\text{j}, \cdots, \text{x}_\text{k}\right\}} \\ = & \frac{\exp[\beta_\text{j}(\text{x}_\text{j}+\text{d})]\exp(\alpha_\text{j})}{\exp(\beta_\text{j}\text{x}_\text{j})\exp(\alpha_\text{j})} \\ = & \exp(\beta_\text{j}\text{d}) \end{aligned} \tag{13}$$

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

Table VII contains such summary statistics for the model (4). The outer quartiles of L_{gnl} and T_{flux} are shown in the columns labelled with "Low" and "High", respectively. So the half-sample odds ratio for L_{gnl} is 5.18, with 0.95 confidence interval [2.22, 12.09], when T_{flux} is set to its median. The effect of increasing L_{gnl} from 7.190 (its lower quartile) to 53.190 (its upper quartile) is to increase the log odds by 1.64 or to increase the odds by a factor of 5.18. The value of odd ratio for T_{flux} is nearly same as L_{gnl}.

Instead of displaying the result in odds, Figure 6 directly shows the predicted probabilities versus each predictive variables (models (1)-(3)). The probability curves for C-, M- and X-class flares are plotted in

black, red and green color, respectively. The plot indicates that: (1) the occurrence probability for each class of solar flares increases with the predictive parameters, (2) for C-class flare predictions, there is a saturation value. The probabilities are nearly 100% when the measure values are larger than their thresholds. For M- and X-class probabilities, when $L_{\rm gnl}$, $E_{\rm diss}$ are used as predictors (panels (a) and (c)), no such saturation value exists. The probabilities keep increasing as predictors increase. However, when $T_{\rm flux}$ 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. (3) The maximum predicted probability of X-class flares is only around $0.3 \sim 0.6$. This may suggest that each single magnetic variable is not sufficient to predict X-class flares.

Finally our fitted regression expression of model (4) is shown as following:

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

$$\begin{aligned}
\hat{\alpha}_1 &= -1.01, \\
\hat{\alpha}_2 &= -2.81, \\
\hat{\alpha}_3 &= -4.81,
\end{aligned}$$

and

$$\begin{aligned}
X\hat{\beta} = & \\
& +5.13\times10^{-2}\,\mathrm{L_{gnl}} \\
& +2.42\times10^{-2}\mathrm{T_{flux}} + 2.84\times10^{-4}(\mathrm{T_{flux}} - 5.23)_+^3 \\
& -9.97\times10^{-4}(\mathrm{T_{flux}} - 16.73)_+^3 + 8.81\times10^{-4}(\mathrm{T_{flux}} - 26.7)_+^3 \\
& -1.68\times10^{-4}(\mathrm{T_{flux}} - 49.6)_+^3 \\
& +\mathrm{L_{gnl}}[3.82\times10^{-5}\,\mathrm{T_{flux}} - 3.73\times10^{-6}(\mathrm{T_{flux}} - 5.23)_+^3 \\
& +1.14\times10^{-5}(\mathrm{T_{flux}} - 16.73)_+^3 - 9.11\times10^{-6}(\mathrm{T_{flux}} - 26.7)_+^3 \\
& +1.46\times10^{-6}(\mathrm{T_{flux}} - 49.6)_+^3]
\end{aligned}$$

and $(x)_+ = x$ when $x > 0$, $(x)_+ = 0$ otherwise.

L_{gnl} and T_{flux}, measured for a given active region, are then put into the above equation to compute the predicted probabilities.

4.4. COMPARISON WITH NOAA/SEC 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 historical rate of number of flares produced by a given sunspots group. This approach is the basis of the prediction generated by NOAA/SEC [4] and NASA Goddard Space Flight Center's Solar Data Analysis Center (SDAC). (Gallagher, Moon and Wang, 2002).[5] In addition to the McIntosh classification scheme, NOAA/SEC 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.

Disadvantages of the classification-based approaches are that the variation in flare probability within a class is unavoidably ignored. The classification process is possibly subjective because the McIntosh scheme with three parameters is an arbitrary construction. Different observers may not agree with a given classification. The 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.

In order to compare the predictability of the Logistic model and NASA/SDAC, NOAA/SEC schemes, we studied our event list and found 55 events in the list were also predicted by NOAA/SEC and NASA/SDAC. Their prediction results were plotted together and shown in Figure 7. Every event (flare) is indexed in x-axis. Y axis represents the predicted probability. The results from different prediction approach are indicated by different shapes of points. For comparison, the actual results (1 means occurred, 0 means not) are also presented (green dots). We then used a contingency table, 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 experience 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 horizontal dotted line. The vertical dotted line indicates the actual start point of flare happening. Each graph in Figure 7 is divided into four regions (a-d). Region "a" contains the events with "yes predicted" and "yes observed". The region "b" represents the number of false alarms that means

[4] `http://www.sec.noaa.gov/ftpdir/latest/daypre.txt`

[5] `http://www.solarmonitor.org`

"yes predicted" but not observed. Similarly, "c" is the number of misses that means not predicted but "yes observed", and "d" is the number of correct nulls that means not predicted or observed.

The indexes used by NOAA National Weather Service (NWS) were computed and listed in Table VIII. POD is the percentage of all flare events which are predicted (a/(a+c). A perfect score would be 100%). FAR measures how often we issue false alarm, or in other words, 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-class flare prediction, the predicted probabilities computed from NASA/SDAC only distribute in the range of 0 and 55%. The "yes predicted" is not as obvious as those from other two methods. Meanwhile, the minimum probabilities predicted by Logistic method are larger than the results from both NASA/SDAC and NOAA/SEC. This is probably due to the threshold (50 GMm^{-1} in this study) for the gradient neutral lines. Those L_{gnl} with small values might still have enough nonpotential energy to product weaker flares. For M- and X-class flares, such a problem is eliminated. In M-class prediction, NASA/SDAC approach is no doubt incapable to satisfy the prediction requirement. For X-class prediction, the results from all current methods are not satisfactory. Thus, the indexes show that the method used by NOAA/SEC provide 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 close enough correlation in triggering stronger flares. The other possible reason for the incapability in prediction of X class flares may be due to the insufficient data samples in logistic regression model.

The gap between NOAA/SEC and logistic regression model become smaller when forecasting major solar flares. In Figure 7 the probabilities of X class flares prediction obtained from ordinal logistic method and NOAA/SEC are higher in those active regions producing flares. We therefore lower the cutoff probability to 25% and recount the value of a, b, c, and d. The resulted indexes are displayed in the last two columns of Table VIII. Every index of logistic method is better that the one from NOAA/SEC. We propose that Ordinal logistic method is more promising in forecast major flares, especially as we have enough data samples, and even more predictive parameters in the future.

5. Conclusions

In this paper we proposed a statistical ordinal logistic regression model to solar flare prediction. For this, we have selected 230 active regions from 1996 to 2005, computed their corresponding magnetic parameters L_{gnl}, T_{flux} and E_{diss} measured from SOHO MDI magnetograms and then applied logistic model to them. Our main results can be summarized as follows.

1. The ordinal logistic regression model is proved to be a viable approach to the automated flare prediction. The results are much better than those data published in NASA/SDAC service, and comparable to the data provided by NOAA/SEC complicated expert system. To our knowledge, this is the first time that logistic regression model is applied in solar physics to predict flare occurrence. And this is the first time that the occurrence probability of flares is quantified into math expression.

2. Each magnetic parameters on photospheric layers $L_{\rm gnl}$, $T_{\rm flux}$ and $E_{\rm diss}$ has a positive correlation with the predicted probability. Among them the most significant variable is $L_{\rm gnl}$, followed by the $T_{\rm flux}$ and $E_{\rm diss}$.

3. Considering the interaction effects between predictive parameters, statistical analysis demonstrates the combination of $L_{\rm gnl}$ and $T_{\rm flux}$ might be enough to be included in the prediction model.

4. According to the results from contingency table, we found that all three approaches can get good results in forecasting C-class flares (CSI is between $0.64 \sim 0.75$). In the M-class prediction, only Logistic and expert system approach are feasible (0.61 and 0.66, respectively). For X-class flare prediction, the 50% cutoff is too strict to all methods to achieve. It perhaps implies that the current parameters used in prediction are not sufficient enough to forecast these super flares. After we changed the cutoff probability to 25%, both methods might be acceptable. However, ordinal logistic method provided better performance and is more promising in X class prediction.

So far our prediction model is limited to those magnetic parameters obtained only through SOHO MDI magnetograms. There are several physical parameters which 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) length of strong-sheared magnetic neutral line (Falconer et al., 2003); (2) total unsigned vertical current $\int J_z dA$, where J_z is the vertical current density, and (3) photospheric excess magnetic energy $\int \rho_e dA$, where ρ_e is the density of the excess magnetic energy (Wang, et al, 1996, Leka and Barnes, 2003a,b). More extensive investigation is in preparation as these parameters become readily available in the near future.

Acknowledgements

SOHO is an international cooperation project of NASA and ESA. The work is supported by NSF under grants IIS-0324816, ATM-0548952, ATM-0342560 and ATM-0536921, NASA under grants NNG0-6GC81G. VY's additional support is from NASA grants NNG0-5GN34G and NASA ACE NNG0-4GJ51G.

References

Abramenko, V.I., Gopasyuk, S.I., and Ogir', M.B.: 1991, *Solar Phys.* **134**, 287.

Abramenko, V.I., Yurchyshyn, V.B., Wang, H., Spirock, T.J., and Goode, P.R.: 2003, *Astrophys. J.* **597**, 1135.

Abramenko, V.I.: 2005, *Astrophys. J.* **629**, 1141.

Agresti, A.: 1996, *An introduction to categorical data analysis*, New York: Wiley.

Alzola, C.F. and Harrell, F.E.:2004, *An Introduction to S and the Hmisc and Design Libraries*, Free available electronic book.

Antalova, A.: 1996, *Contrib. Astron. Obs.*, Skalnate Pleso, 26, 98.

Armstrong, B.G. and Sloan, M.: 1989, *Am. J. Epidemiol.* **129**, 191.

Bornmann, P.L. and Shaw, D.:1994, *Solar Phys.* **150**, 127

Breiman, L.:1992, *Journal of the American Statistical Association 87*, 738.

Brueckner, G.E. *et al.*: 1998, *Geophys. Res. Lett.* **25**, 3019.

Cane, H.V., Richardson, I.G., and St. Cyr, O.C.: 2000, *Geophys. Res. Lett.* **27**, 3591.

Canfield, R.C., Hudson, H.S., and McKenzie, D.E.: 1999, *Geophys. Res. Lett.* **26**, 627.

Cragg, J.G. and Uhler, R.: 1970, *Canadian Journal of Economics*, **3**, 386.

Efron, B.:1986, *Journal of the American Statistical Association* **81**, 461

Falconer, D.A.: 2001, *J. Geophys. Res.* **106**, 25185.

Falconer, D.A., Moore, R.L., and Gary, G.A.: 2003,*J. Geophys. Res.* **108**, SSH11.

Gallagher, P.T., Moon, Y.J., and Wang, H.:2002, *Solar Phys.* **209** 171.

Gopalswamy, N. *et al.*: 2000, *Geophys. Res. Lett.* **27**, 145.

Hagyard, M.J. *et al.*: 1984, *Solar Phys.* **91**, 115.

Harrell, F.E.: 2001, *Regression Modeling Strategies With Application to Linear Models, Logistic Regression, and Survival Analysis*, Springer.

Hess, W.N.: 1964 (ed.), *The Physics of Solar Flares (Washington: NASA)*.

Jing, J., Song, H., Abramenko, V. I., Tan, C. and Wang, H.: 2006, *Astrophys. J.* **644**, 1273.

Kim, R.S., Cho, K.S. *et al.*:2005, *J. Geophys. Res.* **110**, 11104.

Leka, K. D., Canfield, R. C., McClymont, A. N., de La Beaujardiere, J. F., Fan, Y. and Tang, F.: 1993, *Astrophys. J.* **411**, 370L.

Leka, K. D. and Barnes, G.:2003, *Astrophys. J.* **595**, 1277

Leka, K. D. and Barnes, G.:2003, *Astrophys. J.* **595**, 1296

McCullagh, P.: 1980, *J. Royal Statistical Society* **B42**, 109.

McIntosh, P.S.:1990, *Solar Phys.* **125**, 251

Moreton, G.E. and Severny, A.B.: 1968, *Solar Phys.* **3**, 282.

Nagelkerke, N.J.D.: 1991, *Biometrika*, p205,247,493.

Priest, E. and Forbes, T.: 2000, *Magnetic Reconnection: MHD Theory and Applications (Cambridge: Cambridge Univ. Press)*

Sawyer, C., Warwick, J. W. and Dennett, J. T.: 1986, *Solar flare prediction, Boulder: Colorado Assoc. Univ. Press*

Scott, S. C., Goldberg, M. S. and Mayo, N. E.: 1997, *J. Clinical Epidemiology* **50**, 45.

Song, H., Yurchyshyn, V., Yang, G., Tang, C., Chen, W. and Wang, H.: 2006, *Solar Phys.* in press.

Stone, C.J. and Koo, C. Y.: 1985, *Proceeding of the Statistics Computing Section ASA, Washington, DC* 45.

Svestka, Z.: 1976, *Solar Flares (Dordrecht: Reidel)*.

Tian, L., Liu, Y. and Wang, J.: 2002, *Solar Phys.* **209**, 361.

Wang, H.: 2006, *Astrophys. J.* **649**, 490.

Wang, H., Song, H., Yurchyshyn, V., Deng, Y., Zhang, H., Falconer, D., and Li, J.: 2006, *ChJAA* **6**, 477.

Wang, J., Shi, Z., Wang, H. and Lü, Y.: 1996, *Astrophys. J.* **456**, 861.

Wang, T., Xu, A. and Zhang, H.: 1994, *Solar Phys.* **155**, 99.

Wang, Y.M., Ye, P.Z., Wang, S., Zhou, J.P., and Wang, J.: 2002, *J. Geophys. Res.* **107**, 2

Webb, D.F. *et al.*: 2000, *J. Geophys. Res.* **105**, 7491

Zhang, J. *et al.*: 2003, *Astrophys. J.* **582**, 520.

Table I.: List of Active Regions Associated with Flares

Date	AR	Location	L_{gnl} (Mm)	$T_{flux}(10^{21}$ Mx)	$E_{diss}(10^5 Jm^{-1}s^{-1})$	F_{idx}	Level
20050117	0721	S04E03	0.00	4.00	2.41	0.01	0
20050123	0726	N01W00	0.00	6.94	4.40	0.01	0
20050202	0729	S10W09	0.00	5.30	3.86	1.63	0
20050208	0731	S02W01	0.00	3.94	2.04	0.01	0
20050302	0739	S03W03	0.00	4.65	2.66	0.84	0
20050315	0743	S08W03	5.75	13.80	7.64	14.58	0
20050402	0747	S06W04	0.00	7.19	6.07	6.84	0
20050408	0749	S05E11	0.00	5.98	4.27	0.01	0
20050411	0750	S07E08	4.31	11.90	9.53	1.25	0
20050508	0758	S07E08	18.69	21.90	21.40	140.88	0
20050604	0769	S06E01	0.00	12.30	8.70	0.80	0
20050610	0775	N10E06	48.87	17.30	13.30	65.48	0
20050804	0796	S07W01	0.00	5.21	3.11	0.17	0
20050818	0798	S09E08	0.00	5.26	3.32	120.27	0
20051007	0813	S08E01	10.06	10.30	8.76	1.40	0
20051020	0815	N08E07	0.00	5.39	2.74	0.01	0
20051102	0819	S09W05	0.00	6.64	3.37	1.71	0
20051103	0818	S08W04	0.00	6.14	2.99	0.20	0
20051126	0825	S06E01	0.00	4.66	2.04	0.01	0
20051215	0834	S07W01	0.00	12.20	6.92	2.76	0
20051229	0840	S03E02	0.00	9.70	5.59	0.01	0
19980113	8131	S24W12	35.94	13.20	10.40	36.44	0
19990811	8662	S16E08	20.12	21.90	19.00	35.64	0
20010219	9354	S09W07	8.62	13.80	8.40	12.49	0
20010710	9531	S06E05	10.06	12.40	7.05	13.40	0
20010718	9545	N09E03	2.87	8.61	4.24	11.26	0
20010720	9542	N08E07	0.00	6.69	3.08	0.49	0
20010731	9557	S21E25	0.00	10.90	7.11	135.02	0
20020508	9937	S09E13	10.06	12.60	10.00	37.63	0
20020613	9991	S20E05	0.00	14.00	8.67	5.44	0
20020618	0000	N18E15	0.00	14.40	10.70	341.48	0
20021204	0208	N09E03	31.62	16.70	14.00	27.48	0
20030305	0296	N12E05	10.06	23.50	15.70	13.12	0
20030312	0306	N05E06	8.62	21.80	12.20	8.25	0
20030415	0334	S08E12	0.00	10.80	7.30	0.55	0
20030517	0357	S17E07	10.06	6.16	4.49	0.85	0
20030525	0365	S09E21	7.19	10.70	9.20	599.27	0
20030620	0388	S03E04	4.31	8.31	7.22	10.67	0
20030909	0456	S09E10	0.00	8.64	6.76	15.26	0
20031006	0471	S08E07	8.62	23.30	19.90	48.83	0
20040112	0537	N04W04	25.87	15.00	10.00	271.55	0

Continued on next page

Table I – continued from previous page

Date	AR	Location	L_{gnl} (Mm)	$T_{flux}(10^{21}$ Mx)	$E_{diss}(10^5 Jm^{-1}s^{-1})$	F_{idx}	Level
20040224	0564	N14E00	18.69	22.50	20.00	238.04	0
20040518	0617	S12E08	0.00	6.42	4.84	3.24	0
20040525	0618	S10E12	10.06	27.70	21.90	107.47	0
20040603	0621	S14E13	8.62	18.40	12.80	6.35	0
20040606	0624	S08E10	0.00	7.54	4.62	0.34	0
20040804	0655	S09E14	12.94	15.90	13.20	10.98	0
20041002	0675	S10W06	0.00	9.78	7.03	0.65	0
20041023	0684	S03W00	7.19	12.20	10.50	0.61	0
20041125	0704	N13W18	7.19	14.70	11.60	2.12	0
20041201	0706	S08W16	0.00	13.40	11.20	27.57	0
20050215	0735	S08E07	4.31	9.71	7.83	13.93	0
20050312	0742	S05E03	7.19	23.60	19.50	25.79	0
20050418	0754	S08E06	2.87	6.58	4.84	0.40	0
20050507	0758	S09E26	31.62	19.00	17.40	140.88	0
20050611	0776	S06E04	8.62	22.30	15.50	37.30	0
20050726	0791	N14E23	7.19	10.90	9.07	7.81	0
20050815	0797	S13E12	10.06	12.30	9.80	0.89	0
20051102	0818	S08E09	0.00	5.66	3.11	0.20	0
20051126	0824	S14W09	0.00	10.10	5.73	10.09	0
20051204	0828	S04E04	0.00	7.19	3.95	1.34	0
20051215	0835	N19W03	7.19	9.63	5.22	7.13	0
20051219	0837	S10W10	0.00	10.60	7.36	2.04	0
19981104	8375	N19W08	61.81	23.20	14.50	220.89	1
19990602	8562	S16E07	54.62	17.10	15.20	21.70	1
19990626	8598	N23E09	0.00	30.10	24.80	71.20	1
19990629	8603	S15E16	0.00	23.40	18.00	77.20	1
19990701	8611	S25E18	0.00	17.70	15.70	160.70	1
19990802	8651	N24E08	47.44	45.40	30.40	153.10	1
19990803	8651	N25W04	47.44	42.70	29.80	153.10	1
19990826	8674	S22E09	43.12	47.10	36.70	346.70	1
19991111	8759	N09E14	138.00	35.30	26.70	113.50	1
19991125	8778	S15E06	14.37	18.60	15.70	138.90	1
19991127	8778	S14W17	35.94	20.60	15.40	138.90	1
20000316	8910	N11E18	12.94	22.90	16.10	437.51	1
20000410	8948	S15E03	64.69	27.90	21.10	216.10	1
20000418	8963	N16E18	0.00	14.60	8.77	54.70	1
20000419	8963	N14E09	23.00	14.00	10.10	54.70	1
20000517	8996	S20E16	48.87	43.10	33.60	129.40	1
20000608	9026	N20W06	56.06	31.70	31.50	945.23	1
20000707	9070	N20E14	46.00	23.10	24.00	186.80	1
20000708	9070	N17W01	43.12	24.80	26.60	186.80	1
20000905	9154	S20E06	24.44	19.00	18.50	55.56	1

Continued on next page

Table I – continued from previous page

Date	AR	Location	L_{gnl} (Mm)	$T_{flux}(10^{21}$ Mx)	$E_{diss}(10^5 Jm^{-1}s^{-1})$	F_{idx}	Level
20000930	9173	S12E13	34.50	19.20	14.40	50.30	1
20001009	9182	N02W04	0.00	12.70	5.57	69.50	1
20001031	9209	S23W06	20.12	17.20	11.70	81.20	1
20001122	9236	N20E12	10.06	17.60	9.50	1326.30	1
20010306	9368	N26W08	0.00	21.20	15.20	167.00	1
20010327	9393	N18E08	155.25	57.20	47.30	2954.50	1
20010521	9461	N22E08	0.00	16.90	11.30	18.36	1
20010715	9539	S17W01	28.75	11.40	8.46	60.60	1
20010910	9608	S23E14	44.56	37.40	24.90	498.24	1
20010911	9608	S29E10	125.06	35.60	23.90	498.24	1
20010913	9610	S13W08	35.94	36.00	19.10	31.60	1
20010924	9628	S18E07	70.44	38.90	22.80	274.00	1
20010930	9636	N12W05	69.00	27.20	19.20	100.30	1
20011024	9672	S17E00	47.44	28.80	17.00	475.10	1
20011027	9678	N07E05	7.19	28.60	16.60	103.10	1
20011030	9682	N12E02	76.19	41.90	25.10	269.70	1
20011103	9684	N05W17	56.06	24.40	14.80	145.00	1
20011120	9704	S17W09	51.75	26.90	14.70	283.60	1
20020106	9767	S21W14	7.19	31.80	15.60	61.50	1
20020108	9773	N14E05	24.44	26.30	17.30	290.56	1
20020110	9773	N14W17	40.25	34.70	24.40	290.56	1
20020716	0030	N21E01	73.31	44.60	38.70	793.73	1
20020727	0039	S17E17	132.25	55.20	51.10	733.80	1
20020729	0050	S07E06	11.50	21.50	19.30	60.20	1
20020802	0057	S09E05	8.62	9.96	8.25	72.70	1
20020905	0096	N08W01	23.00	27.60	19.80	23.80	1
20021002	0137	S20E18	46.00	15.00	14.90	174.64	1
20021105	0177	N16W09	43.12	23.70	20.00	80.30	1
20021106	0180	S09W07	56.06	25.10	22.80	259.50	1
20030222	0290	N17W06	8.62	15.70	11.80	36.06	1
20030315	0314	S15W13	30.19	14.50	16.60	529.20	1
20030501	0349	S13E07	8.62	34.80	22.50	86.37	1
20030607	0375	N11E09	30.19	26.30	25.30	1358.62	1
20030608	0375	N11W03	43.12	31.80	29.70	1358.62	1
20030718	0410	S12E09	0.00	2.27	0.70	91.71	1
20030815	0431	S13W02	0.00	4.34	2.10	124.65	1
20031028	0488	N09W05	92.00	38.50	46.60	881.80	1
20040225	0564	N14W13	0.00	2.61	0.90	238.04	1
20040329	0582	N13E18	0.00	2.25	0.65	144.65	1
20040331	0582	N13W14	8.62	20.60	13.80	144.65	1
20040719	0649	S09W00	0.00	4.10	2.12	1381.59	1
20040811	0656	S14E13	0.00	3.49	1.54	1260.24	1

Continued on next page

Table I – continued from previous page

Date	AR	Location	L_{gnl} (Mm)	$T_{flux}(10^{21}$ Mx)	$E_{diss}(10^5 Jm^{-1}s^{-1})$	F_{idx}	Level
20050604	0772	S18E09	7.19	10.90	11.10	98.41	1
20050702	0785	S17E04	0.00	3.87	1.86	15.56	1
20050914	0808	S11E02	92.00	44.70	45.60	4886.56	1
19980315	8179	S24W04	53.19	31.40	20.60	100.32	2
19980326	8185	S24E04	17.25	18.20	12.90	48.46	2
19980501	8210	S17E05	7.19	20.00	9.37	422.59	2
19990630	8603	S14W01	18.69	19.10	17.00	77.20	2
19990702	8611	S26E08	33.06	22.60	20.40	160.70	2
19990724	8636	N20W06	33.06	35.00	26.40	94.99	2
19990819	8672	N16W02	10.06	15.10	10.70	10.00	2
19990827	8674	S21W04	74.75	49.60	38.40	346.70	2
19991112	8759	N10E05	90.56	42.90	32.50	113.50	2
19991126	8778	S14W06	21.56	20.20	15.30	138.90	2
19991222	8806	N19E09	24.44	37.40	25.70	259.78	2
20000118	8831	S17E00	7.19	24.90	14.80	49.00	2
20000217	8872	S28E05	0.00	11.30	6.86	13.80	2
20000313	8906	S17E02	107.81	46.30	28.80	284.10	2
20000720	9087	S12W02	34.50	36.70	30.20	443.60	2
20000725	9097	N06W02	30.19	25.60	13.90	149.80	2
20000916	9165	N15E00	27.31	21.70	14.60	259.60	2
20001109	9221	S12E08	0.00	13.70	8.49	10.00	2
20001118	9231	S21E00	15.81	22.10	17.30	99.01	2
20001123	9236	N22E04	58.94	26.30	17.50	1326.30	2
20010110	9302	N19W00	21.56	20.40	10.90	56.10	2
20010328	9393	N17W04	161.00	62.10	56.80	2954.50	2
20010409	9415	S21E04	50.31	33.80	31.30	2811.82	2
20010425	9433	N19E04	35.94	38.60	35.60	541.09	2
20010505	9445	N25W02	57.50	26.70	20.90	70.80	2
20010513	9455	S17E01	48.87	16.10	17.60	161.04	2
20010604	9484	S06E05	20.12	13.20	7.55	37.00	2
20010903	9601	N13E02	73.31	39.20	22.30	327.31	2
20010925	9628	S20E00	120.75	46.60	31.00	274.00	2
20010929	9636	N16E07	17.25	25.90	15.90	100.30	2
20011106	9687	S20E01	37.37	25.40	14.60	333.10	2
20011110	9690	S17E05	156.68	54.10	34.00	518.83	2
20011111	9690	S17W07	136.56	46.00	25.90	518.83	2
20011129	9715	N04E03	69.00	36.40	22.10	262.60	2
20020109	9773	N14W04	38.81	34.60	26.00	290.56	2
20020314	9866	S09E06	11.50	32.50	22.10	163.70	2
20020315	9866	S09W06	15.81	32.10	24.50	163.70	2
20020410	9893	N19W08	10.06	19.40	14.20	248.70	2
20020415	9906	S14W04	51.75	30.40	20.80	215.82	2

Continued on next page

Table I – continued from previous page

Date	AR	Location	L_{gnl} (Mm)	$T_{flux}(10^{21}$ Mx)	$E_{diss}(10^5 Jm^{-1}s^{-1})$	F_{idx}	Level
20020728	0039	S16E08	143.75	52.40	46.70	733.80	2
20020728	0044	S18E01	70.44	48.40	41.60	309.70	2
20020815	0066	N13E03	8.62	17.50	14.10	22.40	2
20020817	0069	S08E08	169.62	56.60	47.70	1100.00	2
20020818	0069	S08W07	173.93	59.90	46.00	1100.00	2
20020823	0083	S18W05	50.31	16.00	17.30	135.80	2
20021003	0137	S19E08	44.56	17.90	17.40	174.64	2
20021004	0137	S19W05	14.37	18.90	18.20	174.64	2
20021025	0162	N27W03	44.56	33.80	26.00	246.48	2
20021216	0227	N06W06	8.62	9.31	10.40	28.30	2
20021217	0226	S27W02	48.87	27.80	28.50	231.60	2
20021219	0229	N19W02	0.00	25.90	20.50	42.30	2
20030107	0244	S21W01	0.00	14.40	11.00	40.17	2
20030123	0266	N13W04	12.94	8.28	7.51	65.81	2
20030421	0338	N18E06	0.00	8.66	6.13	399.41	2
20031024	0484	N02E01	81.94	49.60	34.60	696.70	2
20031118	0501	N01E08	64.69	22.30	16.00	404.78	2
20031119	0501	N01W03	47.44	21.20	15.00	404.78	2
20040118	0540	S14E01	38.81	26.60	19.90	179.69	2
20040723	0652	N08E04	66.12	57.10	42.40	670.64	2
20040812	0656	S13E02	27.31	39.60	36.80	1260.24	2
20041105	0696	N09E06	84.81	26.20	22.70	1120.55	2
20041106	0696	N09W08	80.50	30.10	31.30	1120.55	2
20041202	0708	N09E01	0.00	13.10	9.24	31.34	2
20050114	0718	S07W08	40.25	19.20	19.60	87.67	2
20050517	0763	S17E06	41.69	14.00	16.80	130.91	2
20050707	0786	N11E08	51.75	20.70	22.50	612.87	2
20051118	0822	S08W01	17.25	23.60	10.90	255.59	2
20051202	0826	S04E06	21.56	22.70	16.80	221.05	2
19980502	8210	S17W12	37.37	23.10	12.40	422.59	3
20000606	9026	N21E18	51.75	30.50	26.20	945.23	3
20000607	9026	N20E05	53.19	29.80	31.00	945.23	3
20000711	9077	N17E45	109.25	27.70	19.50	1256.40	3
20000712	9077	N18E27	92.00	35.80	27.60	1256.40	3
20000714	9077	N17E02	76.19	37.30	38.00	1256.40	3
20001124	9236	N21W10	74.75	27.80	17.50	1326.30	3
20001125	9236	N21W24	37.37	26.70	16.20	1326.30	3
20010329	9393	N17W18	182.56	59.40	59.50	2954.50	3
20010406	9415	S21E42	100.62	21.90	19.20	2811.82	3
20010410	9415	S22W12	84.81	32.00	30.10	2811.82	3
20010623	9511	N10E23	15.81	9.74	7.14	276.79	3
20010825	9591	S18E40	80.50	25.10	17.20	872.30	3

Continued on next page

Table I – continued from previous page

Date	AR	Location	L_{gnl} (Mm)	$T_{flux}(10^{21}$ Mx)	$E_{diss}(10^5 Jm^{-1}s^{-1})$	F_{idx}	Level
20010924	9632	S18E28	51.75	31.00	18.90	322.40	3
20011022	9672	S19E23	44.56	28.00	16.80	475.10	3
20011025	9672	S19W16	58.94	35.90	18.30	475.10	3
20011104	9684	N05W29	12.94	22.90	12.50	145.00	3
20020715	0030	N19E11	86.25	39.50	36.60	793.73	3
20030527	0365	S06W08	51.75	23.00	21.10	599.27	3
20031026	0486	S16E41	182.56	44.30	33.20	6829.50	3
20031028	0486	S18E04	240.06	70.60	68.60	6829.50	3
20031029	0486	S17W09	222.81	69.30	58.10	6829.50	3
20040226	0564	N14W28	34.50	28.50	22.90	238.04	3
20040715	0649	S10E48	79.06	24.20	18.80	1381.59	3
20040716	0649	S08E38	63.25	26.50	23.00	1381.59	3
20040717	0649	S08E24	38.81	28.80	26.30	1381.59	3
20040813	0656	S13W12	58.94	43.30	35.60	1260.24	3
20041030	0691	N13W14	24.44	17.70	17.20	454.48	3
20041107	0696	N08W21	64.69	27.90	27.50	1120.55	3
20050101	0715	N04E22	15.81	13.50	12.40	158.56	3
20050115	0720	N13W03	119.31	45.90	36.50	2379.42	3
20050117	0720	N13W29	100.62	39.10	28.70	2379.42	3
20050913	0808	S11E17	130.81	41.60	39.80	4886.56	3
20050915	0808	S11W13	81.94	41.60	41.50	4886.56	3

Table II. Descriptive statistics of solar flares data

	X-class (n=34)		M-class (n=68)		C-class (n=65)		N-class (n=63)	
Label	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
$L_{gnl}(\mathrm{Mm})$	81.18	55.62	47.86	43.72	36.62	36.35	6.75	10.03
$T_{flux}(10^{21}\mathrm{Mx})$	33.23	13.55	29.05	13.63	24.89	13.06	11.91	5.93
$E_{diss}(10^5\,\mathrm{J\,m^{-1}s^{-1}})$	27.52	14.03	22.20	11.12	19.07	11.47	8.69	5.27

Table III. Regression models for different combination of predictive parameters

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

Table IV. Indexes to evaluate the predictive ability of models

Models	R_N^2	c	D_{xy}
(1)	0.382	0.771	0.543
(2)	0.341	0.748	0.496
(3)	0.333	0.749	0.497
(4)	0.432	0.791	0.582
(5)	0.353	0.758	0.516
(6)	0.400	0.782	0.564
(7)	0.430	0.792	0.584
(8)	0.423	0.785	0.569

Table V. Validation of Model With Predictive Variables L_{gnl} and T_{flux}

	index.orig	optimism	index.corrected
D_{xy}	0.579	0.020	0.559
R^2	0.432	0.033	0.399
Intercept	0.000	−0.009	0.009
Slope	1.000	0.067	0.933
E_{max}	0.000	0.017	0.017

Table VI. Validation Results of All Models

Models	Bias-corrected D_{xy}	Bias-corrected R^2	Intercept	Slope	E_{max}
(1)	0.538	0.365	−0.011	0.969	0.009
(2)	0.490	0.325	0.001	0.970	0.007
(3)	0.501	0.326	0.002	0.984	0.004
(4)	0.559	0.399	0.009	0.933	0.017
(5)	0.489	0.309	0.021	0.899	0.027
(6)	0.533	0.362	0.000	0.924	0.018
(7)	0.557	0.382	−0.022	0.898	0.028
(8)	0.551	0.389	0.000	0.928	0.017

Table VII. Effects of L_{gnl}, T_{flux} on response variable *Level*

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
L_{gnl}	7.190	53.190	46.00	1.64	0.43	0.80	2.49
Odds Ratio	7.190	53.190	46.00	5.18		2.22	12.09
T_{flux}	13.125	31.775	18.65	1.61	0.51	0.61	2.62
Odds Ratio	13.125	31.775	18.65	5.03		1.85	13.68

Table VIII. Comparison between three prediction approaches

	C-class flares prediction			M-class flares prediction			X-class flares prediction[1]	
	Logistic	NOAA/SEC	NASA/SDAC	Logistic	NOAA/SEC	NASA/SDAC	Logistic	NOAA/SEC
a: yes predicted	16	18	14	11	12	2	5	4
b: false alarms	5	5	2	2	2	1	2	2
c: misses	4	1	6	5	4	14	2	3
d: correct nulls	30	31	33	37	37	38	46	46
POD: a/(a+c)	0.80	0.95	0.70	0.69	0.75	0.13	0.71	0.57
FAR: b/(a+b)	0.24	0.22	0.13	0.15	0.14	0.33	0.29	0.33
CSI: a/(a+b+c)	0.64	0.75	0.64	0.61	0.67	0.12	0.56	0.44

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

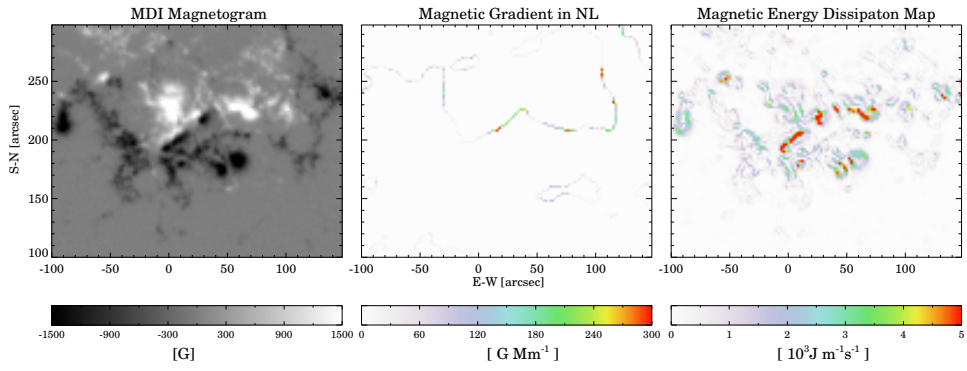

Figure 1. *Left* : line-of-sight magnetogram of NOAA AR 9077 taken on 2000 July 14. *Middle* : Gradient distribution along the neutral line. *Right* : Map of the energy dissipation. The magnitude of parameters in each pixel is indicated by the corresponding color scale bar.

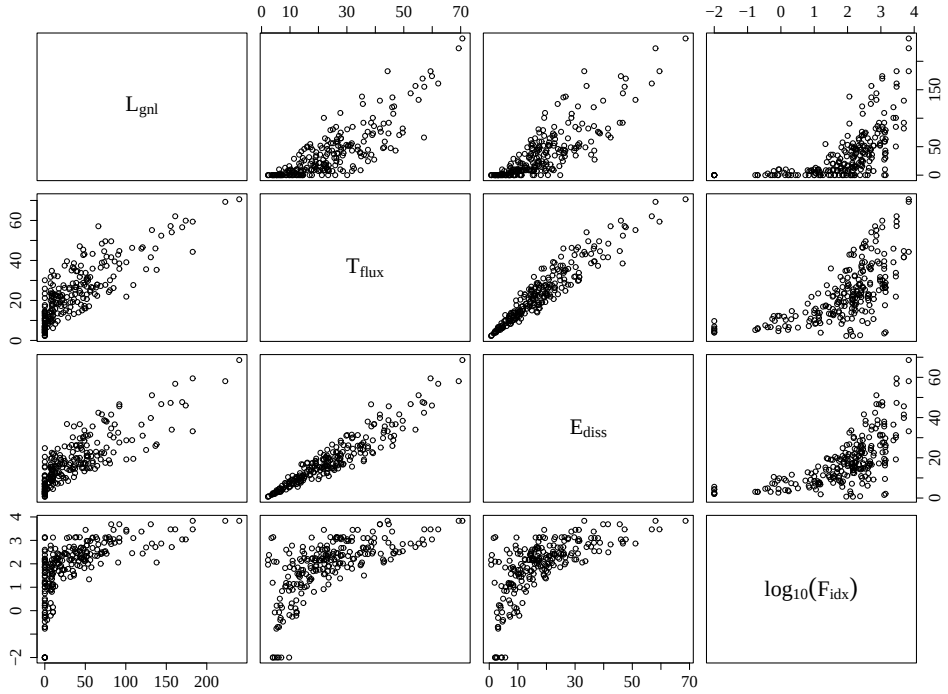

Figure 2. Scatterplots Matrix for L_{gnl}, T_{flux},E_{diss} and F_{idx}. The best correlation is between T_{flux} and E_{diss} (CC is up to 0.95). E_{diss} is the most correlated with flare index among three parameters.

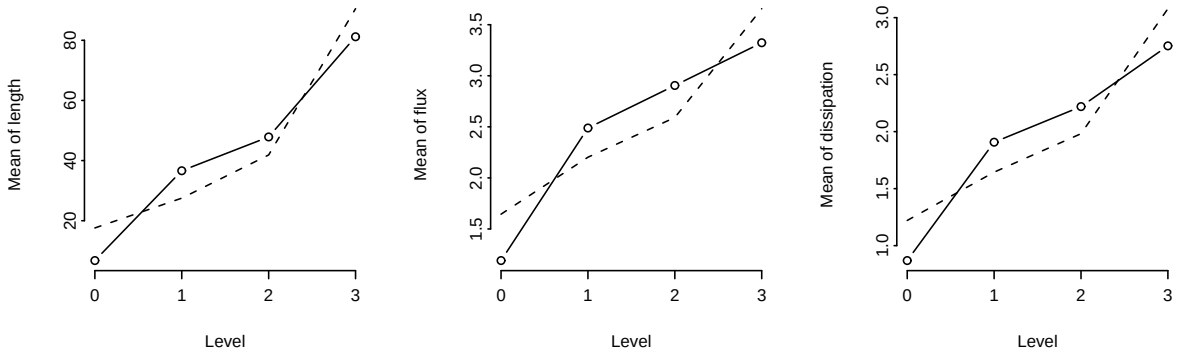

Figure 3. Examination of the ordinality of *Level* for every magnetic parameter by accessing how *Level* relate to the mean value of each predictor, and whether the trend in each plot is monotonic. Solid lines connect the simple stratified means, and dashed lines connect the estimated expected value of X|Y=j given that PO holds. The extend of closeness of two curves indicates the perfect condition to hold ordinal condition.

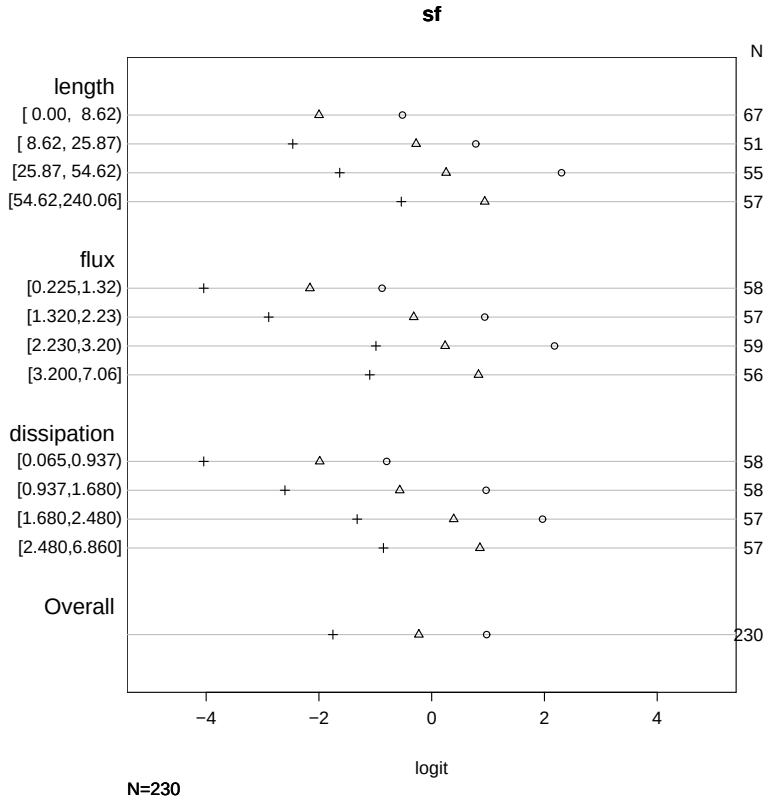

Figure 4. Checking PO assumptions separately for a series of predictive parameters. The circle, triangle, and plus sign correspond to Level $\geq$ 1, 2, 3, respectively. PO is checked by examining the vertical constancy of distances between any two of these three symbols.

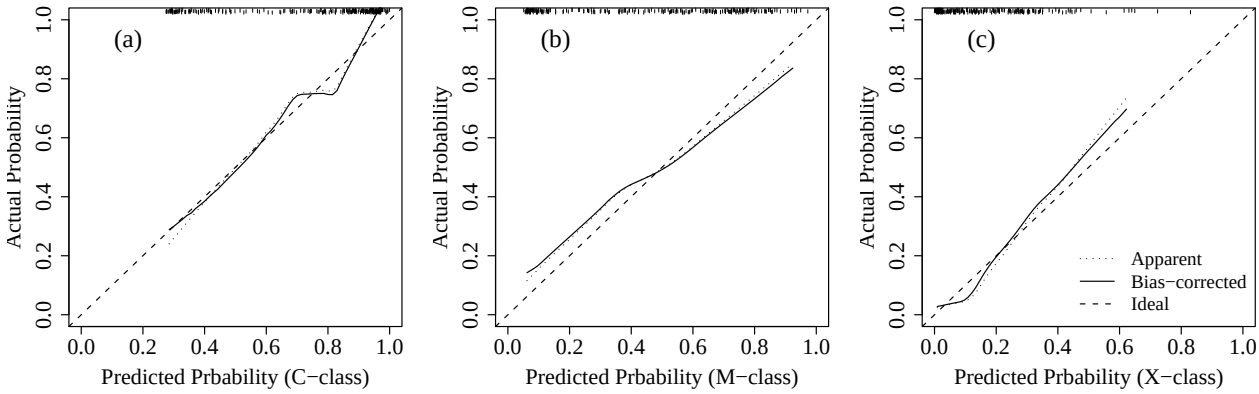

Figure 5. Estimated logistic calibration curves obtained by bootstrapping using the corrected intercept and slope. The logistic calibration model $P_c = [1 + exp(-(\gamma_0 + \gamma_1 L))]$, where P_c is the bias-corrected probability. L is logit($\hat{P}$), and $\hat{P}$ is the predicted probabilities (labelled with 'Apparent'). The bisector line demonstrates excellent validation on an absolute probability scale.

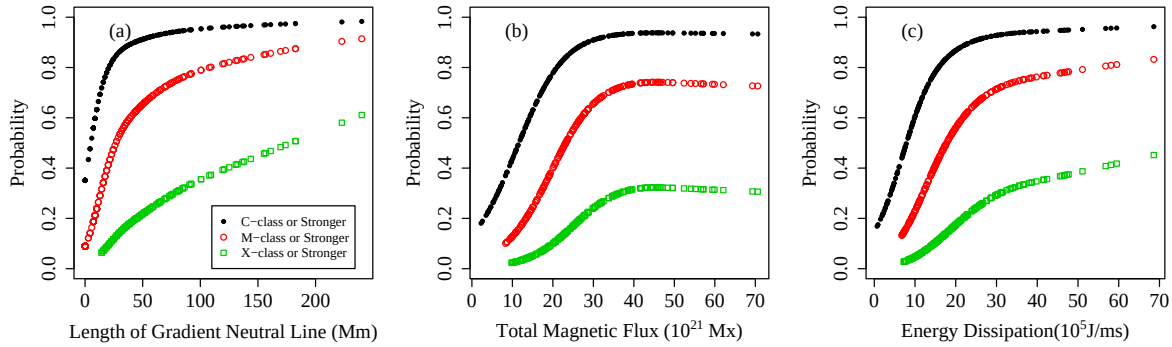

Figure 6. Distribution of predicted occurrence probability of solar flares. Panel (a), (b) and (c) show the results when only L_{gnl}, T_{flux} and E_{diss} as the predictive parameter, respectively. The probabilities for C, M and X class flares are displayed by the black dots, red circles and green squares.

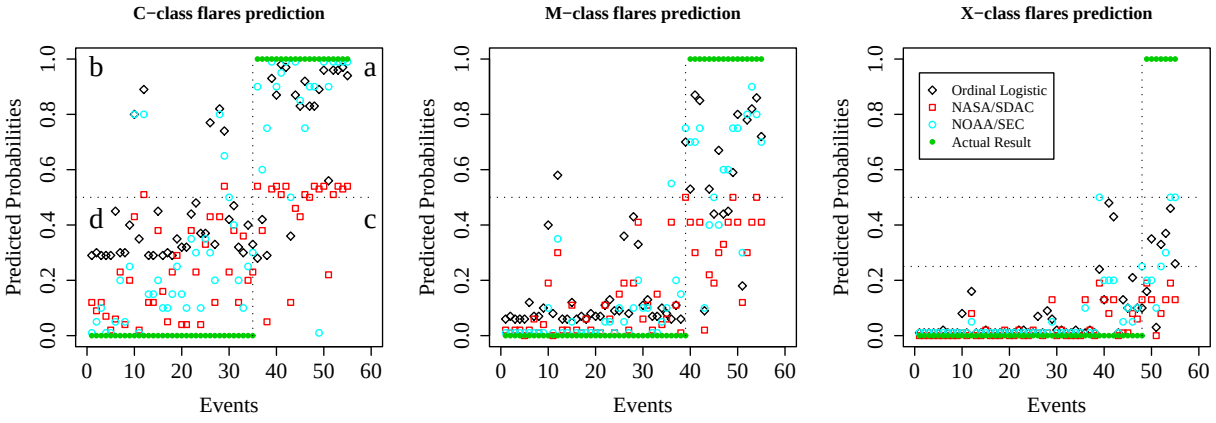

Figure 7. Comparison of three predictive methods for each level of solar flares. The results from Ordinal logistic method, NASA/SDAC and NOAA/SEC are indicated by black diamonds, red squares and blue circles, respectively. For comparison reasons, the actual probabilities of producing flares are shown by green dots. The horizontal dot line is the probability of 50% (One more 25% in X class panel). Vertical dot line represents the turning point of flare occurrence.