Research in Astron. Astrophys. **2012** Vol. **12** No. **3**, 313–321
http://www.raa-journal.org *http://www.iop.org/journals/raa*

Research in Astronomy and Astrophysics

Ensemble prediction model of solar proton events associated with solar flares and coronal mass ejections *

Xin Huang, Hua-Ning Wang and Le-Ping Li

Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; *xhuang@bao.ac.cn*

Received 2011 September 6; accepted 2011 December 13

Abstract An ensemble prediction model of solar proton events (SPEs), combining the information of solar flares and coronal mass ejections (CMEs), is built. In this model, solar flares are parameterized by the peak flux, the duration and the longitude. In addition, CMEs are parameterized by the width, the speed and the measurement position angle. The importance of each parameter for the occurrence of SPEs is estimated by the information gain ratio. We find that the CME width and speed are more informative than the flare's peak flux and duration. As the physical mechanism of SPEs is not very clear, a hidden naive Bayes approach, which is a probability-based calculation method from the field of machine learning, is used to build the prediction model from the observational data. As is known, SPEs originate from solar flares and/or shock waves associated with CMEs. Hence, we first build two base prediction models using the properties of solar flares and CMEs, respectively. Then the outputs of these models are combined to generate the ensemble prediction model of SPEs. The ensemble prediction model incorporating the complementary information of solar flares and CMEs achieves better performance than each base prediction model taken separately.

Key words: solar proton events — Sun: flares — Sun: coronal mass ejections — methods: statistical — ensemble learning

1 INTRODUCTION

Solar proton events (SPEs) mean that the flux of >10 MeV protons exceeds 10 pfu (particles $\mathrm{cm}^{-2}\ \mathrm{s}^{-1}\ \mathrm{sr}^{-1}$). They are harmful to the safety of satellites, the health of astronauts, the quality of communications and so on. Therefore, the prediction of SPEs is important in space weather forecasting services.

The operating prediction models are built by the statistical relationships between SPEs and their precursory phenomena. Among which, the SPE prediction model operated at NOAA's Space Weather Prediction Center (SWPC) was introduced by Balch (1999), and its performance was evaluated by Balch (2008). The inputs of this model are the time-integrated soft X-ray flux, the peak soft X-ray flux, the status of type II and type IV radio bursts, and the location of the associated flare. Kahler et al. (2007) evaluated the performance of the proton prediction system (PPS) developed by Smart & Shea (1989), which predicts the occurrence of SPEs by the properties of solar

∗ Supported by the National Natural Science Foundation of China.

flares. Comparing with the SWPC's prediction model of SPEs, the PPS applies a different modeling algorithm and does not use the type II and type IV radio bursts as inputs. Kubo & Akioka (2004) studied the relationships between SPEs and the soft X-ray flux, and found the threshold for occurrence of SPEs in the parameter space of solar flare duration and peak soft X-ray flux. Garcia (2004) built a logistic regression model of SPEs by the temperature and soft X-ray flux of the associated solar flare, and Laurenza et al. (2009) developed an SPE prediction model by using flare longitude, time-integrated soft X-ray intensity, and time-integrated intensity of type III radio emission near 1 MHz. More recently, Núñez (2011) adopted a dual-model system consisting of the well-connected prediction model and the poorly-connected prediction model for SPE prediction. Moreover, with the development of the field of machine learning, neural networks have been used to build the SPE prediction model from the observational data (Wang 2000; Gabriel & Patrick 2003; Gong et al. 2004). These SPE prediction models mainly depend on the properties of solar flares. However, SPEs are considered to originate from solar flares and/or shock waves associated with coronal mass ejections (CMEs) (Reames 1999). Therefore, CMEs can also provide good indexes for the production of SPEs (Kocharov & Torsti 2002; Gopalswamy et al. 2002, 2008; Ohki 2003; Kahler & Vourlidas 2005; Marqué et al. 2006; Lehtinen et al. 2008; Gerontidou et al. 2009).

Using the machine learning method, we first build two prediction models for SPEs based on the properties of the associated solar flares and CMEs respectively, and then combine these models to form an ensemble model, which fuses the information of solar flares and CMEs. Finally, the performance of the proposed prediction model is evaluated and compared.

The paper is organized as follows: the data set is introduced in Section 2, the modeling algorithm is described in Section 3, the performance of the ensemble prediction model is evaluated in Section 4 and conclusions are given in Section 5.

2 DATA SELECTION

The dataset consists of SPEs and control events, which meet the necessary conditions of SPEs but are not associated with any particular SPE (Balch 2008).

(1) SPEs. This list of SPEs collected by SWPC can be obtained from *http://www.swpc.noaa.gov/ftpdir/indices/SPE.txt*. In it, the majority of SPEs are associated with both solar flares and CMEs. There are two major physical processes accelerating the particles near the Sun. One is related to solar flares, and the other is related to shock waves driven by CMEs. Seventy SPEs associated with both solar flares and CMEs are selected from 1996 (CMEs observed by LASCO began in 1996) to 2005.

(2) Control events. The solar flares (peak flux $\geq$ M1.0) associated with CMEs are considered as the conditions to select control events. Yashiro et al. (2006) studied the relationship between solar flares and CMEs and provided a list of solar flares with CMEs from 1996 to 2005 (*http://cdaw.gsfc.nasa.gov/pub/yashiro/flare_cme/fclist_pub.txt*). From this list, 619 solar flares and CMEs are selected as the control events.

Solar flares are parameterized by their peak flux, duration and longitude, and CMEs are parameterized by their width, speed and measurement position angle (MPA). These parameters are extracted from *ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SOLAR_FLARES/FLARES_XRAY/* for solar flares and *http://cdaw.gsfc.nasa.gov/CME_list/UNIVERSAL/text_ver/univ_all.txt* for CMEs.

The information gain ratio (GR) defined in Equation (1) is used to measure the importance of a parameter to the SPEs

$$GR(y, X_i) = \frac{H(y) - H(y|X_i)}{H(X_i)}, \tag{1}$$

where y is the status of the SPE, X_i stands for the i-th parameter of the precursory phenomena ($i = 1, ..., 6$ represents the flare peak flux, the flare duration, the flare longitude, the CME width,

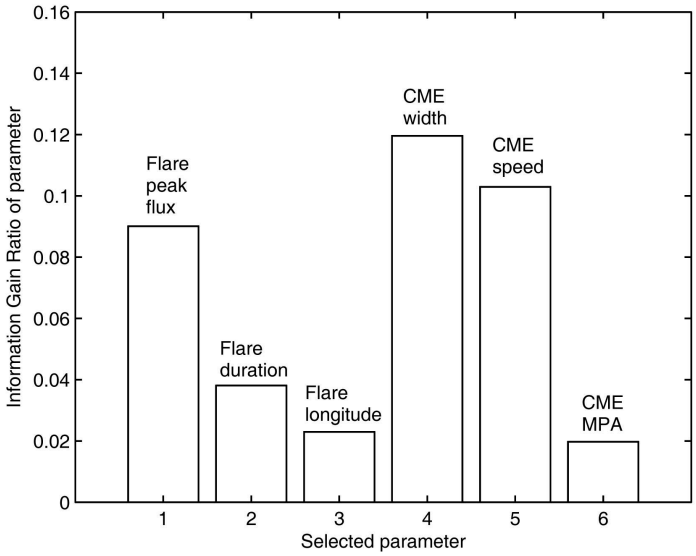

Fig. 1 Information gain ratio for each selected parameter of solar flares and CMEs.

the CME speed and the CME MPA, respectively), and H stands for the information entropy which describes the uncertainty of the system. Here the decrease of uncertainty is called information gain (see the details in Section 3.2 of Yu et al. 2010). Figure 1 shows the information gain ratio for each parameter. From this figure, we find that the CMEs' width and the CMEs' speed are more informative than the flare peak flux and the flare duration for the SPEs' prediction, so it is necessary to consider the information about the CMEs in the SPE prediction model.

3 ENSEMBLE PREDICTION MODEL

It is generally accepted that SPEs are caused by solar flares and/or shock waves associated with CMEs. Therefore two base prediction models of SPEs are built by the properties of solar flares and CMEs, respectively. Then we combine the outputs of these two models to generate the ensemble prediction model whose main modules are shown in Figure 2.

3.1 Base Prediction Model

The hidden naive Bayes (HNB) is used to build the base prediction model from the observational data (Jiang et al. 2009). It is based on the Bayes formula

$$P(y|\mathbf{X}) = \frac{P(\mathbf{X}|y)P(y)}{P(\mathbf{X})}\,, \tag{2}$$

where y is the decision (occurrence or nonoccurrence of SPEs), and $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ is the set of parameters (the properties of solar flares and CMEs).

According to the chain rule of probability, the joint distribution $P(\mathbf{X}|y)$ can be factored as

$$P(\mathbf{X}|y) = \prod_{i=1}^{n} P(X_i|y, X_1, X_2, \cdots, X_{i-1})\,. \tag{3}$$

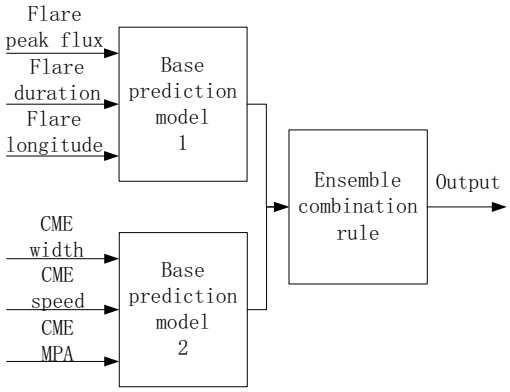

Fig. 2 Main modules for ensemble prediction model of SPEs.

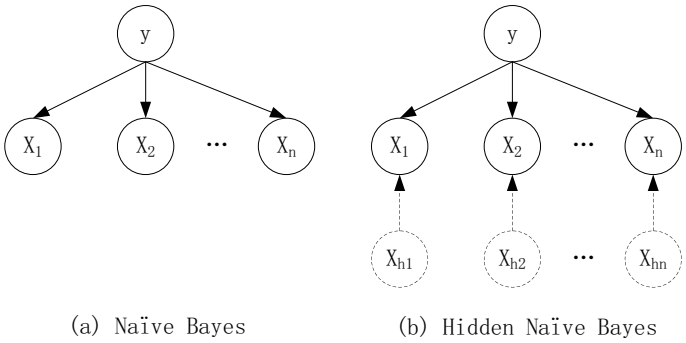

Fig. 3 Structures of naive Bayes (a) and hidden naive Bayes (b). Here y is the status of SPEs, X_i ($i = 1, ..., n$) is the parameter of precursory phenomena, and $X_{\mathrm{h}i}$ is the hidden node of X_i. The arcs between nodes represent the probabilistic dependent relationships. The probability of a node is only dependent on its parent nodes.

Using the conditional independence assumption, that all parameters are independent given the decision y, Equation (3) is simplified as

$$P(\mathbf{X}|y) = \prod_{i=1}^{n} P(X_i|y)\,. \tag{4}$$

The corresponding model is called a naive Bayes model whose structure is shown in Figure 3(a). However, the conditional independence assumption in the naive Bayes case is not true for the prediction of SPEs. Therefore, the hidden nodes ($X_{\mathrm{h}i}$) combine the influences from all other parameters that are created for each parameter (X_i). This improved model is called the hidden naive Bayes method, and its structure is shown in Figure 3(b).

In the hidden naive Bayes model, the joint distribution $P(\mathbf{X}|y)$ is calculated as

$$P(\mathbf{X}|y) = \prod_{i=1}^{n} P(X_i|y, X_{\mathrm{h}i})\,, \tag{5}$$

where

$$P(X_i|y, X_{\mathrm{h}i}) = \sum_{j=1, j\neq i}^{n} W_{ij} \times P(X_i|y, X_j)\,. \tag{6}$$

As shown in Equation (6), the influences from all other parameters are considered, and the weight W_{ij} is the importance of the parameter j compared to the parameter i

$$W_{ij} = \frac{I_P(X_i; X_j|y)}{\sum\limits_{j=1, j\neq i}^{n} I_P(X_i; X_j|y)}\,, \tag{7}$$

where

$$I_P(X_i; X_j|y) = \sum_{x_i, x_j, y} P(x_i, x_j, y) \log \frac{P(x_i, x_j|y)}{P(x_i|y)P(x_j|y)}\,. \tag{8}$$

3.2 Ensemble Combination Rule

The maximum rule (Kittler et al. 1998) is used to combine the outputs of these two base prediction models

$$\mu_j(\mathbf{X}) = \max_{t=1,2} \{P_t(y_j|\mathbf{X})\}\,, \tag{9}$$

where $P_t(y_j|\mathbf{X})$ is the probability of the t-th model for the j-th decision.

When the probabilities of SPE occurrence and nonoccurrence are estimated, a decision threshold can be formulated to provide a binary prediction (whether SPEs will or will not occur).

4 MODEL VALIDATION

4.1 Validation Measures

SPE prediction is considered to be a yes/no prediction in which SPEs do or do not occur. Four possible outcomes of the prediction model are shown in Table 1 (Jolliffe & Stephenson 2003).

Table 1 Yes/No Prediction

	Predicted Yes	Predicted No
Observed Yes	Hit	Miss
Observed No	False alarm	Correct rejection

The numbers of observations in each category (Hit, Miss, False alarm or Correct rejection) are represented by N_{hit}, N_{miss}, N_{FA} and N_{CR}, respectively. The following four validation measures (Jolliffe & Stephenson 2003) are used to estimate the performances of the prediction model.

(1) Hit rate (simply H, which is also called the probability of detection (POD))

$$\mathrm{H} = \frac{N_{\mathrm{hit}}}{N_{\mathrm{hit}} + N_{\mathrm{miss}}}\,. \tag{10}$$

(2) False alarm rate (simply F, which is also called the probability of false detection)

$$\mathrm{F} = \frac{N_{\mathrm{FA}}}{N_{\mathrm{CR}} + N_{\mathrm{FA}}}\,. \tag{11}$$

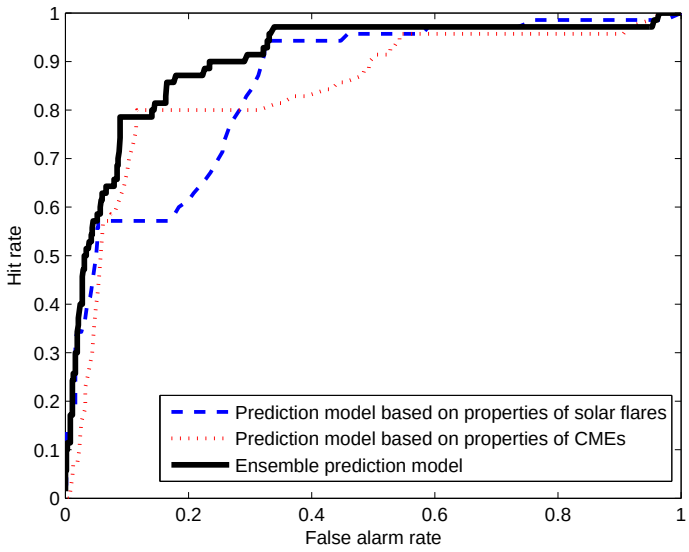

Fig. 4 ROC curves for SPE prediction models.

(3) False alarm ratio (simply FAR)

$$\mathrm{FAR} = \frac{N_{\mathrm{FA}}}{N_{\mathrm{FA}} + N_{\mathrm{hit}}} \,. \tag{12}$$

(4) The Heidke Skill Score (simply HSS)

$$\mathrm{HSS} = \frac{\frac{N_{\mathrm{hit}}+N_{\mathrm{CR}}}{N_{\mathrm{total}}} - \frac{(N_{\mathrm{hit}}+N_{\mathrm{FA}})*(N_{\mathrm{hit}}+N_{\mathrm{miss}})+(N_{\mathrm{CR}}+N_{\mathrm{FA}})*(N_{\mathrm{CR}}+N_{\mathrm{miss}})}{N_{\mathrm{total}}^2}}{1 - \frac{(N_{\mathrm{hit}}+N_{\mathrm{FA}})*(N_{\mathrm{hit}}+N_{\mathrm{miss}})+(N_{\mathrm{CR}}+N_{\mathrm{FA}})*(N_{\mathrm{CR}}+N_{\mathrm{miss}})}{N_{\mathrm{total}}^2}} \,, \tag{13}$$

where

$$N_{\mathrm{total}} = N_{\mathrm{FA}} + N_{\mathrm{hit}} + N_{\mathrm{CR}} + N_{\mathrm{miss}} \,. \tag{14}$$

4.2 Validation Results

The dataset is divided into 10 folds in which nine folds are used to build the prediction model and the remaining one fold is applied to test the model. This process is repeated 10 times until all the data are tested. A single set of binary predictions just shows the performance of a prediction model at a single decision threshold. However, a complete evaluation for the performance of a prediction model requires evaluating the performance over the full range of possible thresholds. Hence, the receiver operating characteristic (ROC) curve (Fawcett 2004), which is a graph of the hit rate (Y-axis) against the false alarm rate (X-axis) for different decision thresholds, is used to estimate the performance of the prediction model.

The ROC curves for three different SPE prediction models are shown in Figure 4. We find that the performances of the prediction models based on the properties of solar flares or CMEs are complementary, and the ensemble SPE prediction model, which fuses complementary information of both solar flares and CMEs, obtains the best performance over the full range of thresholds. The physical explanations for the results are as follows. There are two types of SPEs: impulsive and gradual (Reames 1999; Park et al. 2010). Impulsive SPEs are associated with solar flares and gradual

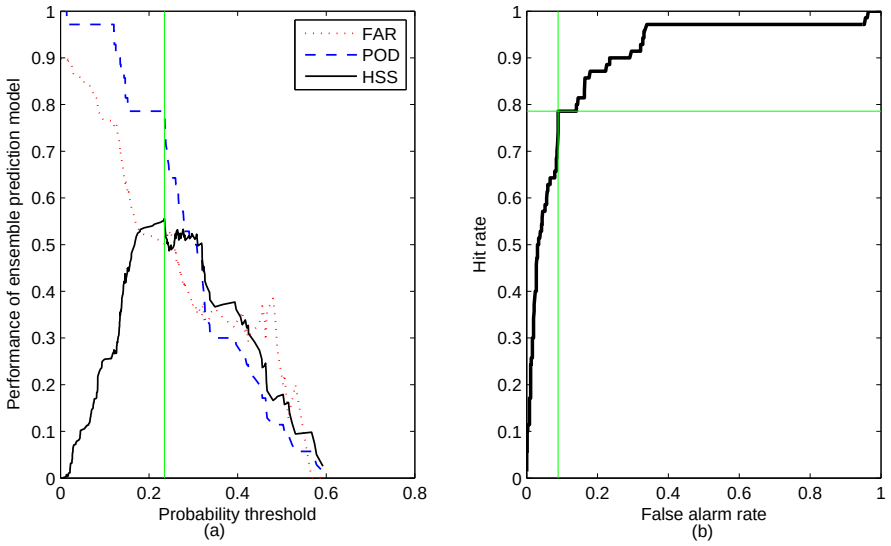

Fig. 5 Performances of the ensemble SPE prediction model. (a) shows performances of the ensemble SPE prediction model traversing the threshold, and the green vertical line marks the optimal threshold for HSS. (b) shows the ROC curve of the ensemble SPE prediction model, and the green horizontal and vertical lines mark the optimal hit rate and false alarm rate, respectively.

SPEs are associated with shocks from CMEs. Although solar flares are correlated with CMEs, the prediction model based on properties of solar flares (dashed line in Fig. 4) is more effective for impulsive SPEs, and the prediction model based on properties of CMEs (dotted line in Fig. 4) is more effective for gradual SPEs. The ensemble prediction model combines the information of these two prediction models, so it obtains good performance for SPE prediction.

Balch (2008) described the commonly used validation measures for the SPE prediction, and completely evaluated the performance of SWPC's prediction model. The SWPC's prediction model of SPEs mainly depends on the properties of solar flares. There are 127 SPEs and 3656 control events from 1986 to 2004 in the dataset. The performance of SWPC's prediction model is evaluated by POD, FAR and HSS, which are functions of a probability decision threshold. A similar figure for the ensemble SPE prediction model is shown in Figure 5(a). For the POD curve, $N_{\rm hit} + N_{\rm miss}$ is a constant and $N_{\rm hit}$ is monotonic with the variation of the threshold, so there is no zigzag in this curve. For the HSS curve, the HSS measures the improvement of the prediction over the standard prediction, and it is not monotonic with the variation of the threshold. Furthermore, the maximum rule is used to combine the outputs of these two base prediction models, and it is not a stable rule. This means that the prediction model based on the properties of solar flares works for some points in the curve, but the prediction model based on the properties of CMEs works for other points in the curve. Hence, some zigzags appear in the HSS curve. Using the HSS as the optimal goal, we notice that the optimal probability threshold is 23.47%, and show the corresponding optimal values in the ROC space (Fig. 5 (b)). At this optimal point, the outputs of the ensemble prediction model are shown in Table 2.

The proposed ensemble prediction model contains the information of CMEs, hence the selection of control events is different between the ensemble prediction model and the SWPC prediction model. Therefore, it is difficult to compare the performance of these two prediction models for the

Table 2 Ensemble SPE Prediction Model at the Optimal Point

	Predicted Yes	Predicted No
Observed Yes	55	15
Observed No	55	564

different testing data. In any case, our conclusion is self-consistent. It means that the performance of the ensemble prediction model is better than that of the prediction model only depending on the properties of solar flares.

5 CONCLUSIONS

Solar flares and CMEs are considered to be two important precursors of SPEs. Hence two base prediction models derived from the properties of solar flares and CMEs are built by HNB, which is a probability-based calculation method from the field of machine learning. The outputs of these two base models are combined by the maximum rule to generate an ensemble prediction model of SPEs. The ensemble model provides a good way to fuse the information about solar flares and CMEs for SPE prediction, and it obtains better performance than each base prediction model taken separately.

The present work focuses on whether solar eruptions will produce SPEs. In the future, peak flux and rise times of SPEs should be estimated by regression models.

Acknowledgements We thank the *SOHO* consortium for the data. *SOHO* is a project of international cooperation between ESA and NASA. This work is supported by the Young Researcher Grant of National Astronomical Observatories, Chinese Academy of Sciences, the National Basic Research Program of China (973 Program, Grant No. 2011CB811406) and the National Natural Science Foundation of China (Grant Nos. 10733020, 10921303, 11003026 and 11078010). This paper has benefited from comments of the reviewer.

References

Balch, C. C. 1999, Radiation Measurements, 30, 231
Balch, C. C. 2008, Space Weather, 6, S01001
Fawcett, T. 2004, ROC graphs: notes and practical considerations for researchers, Technical report (USA:Palo Alto, HP Laboratories)
Gabriel, S. B., & Patrick, G. J. 2003, Space Sci. Rev., 107, 55
Garcia, H. A. 2004, Space Weather, 2, S02002
Gerontidou, M., Mavromichalaki, H., Belov, A., & Kurt, V. 2009, Advances in Space Research, 43, 687
Gong, J.-C., Xue, B.-S., Liu, S.-Q., et al. 2004, Chin. Astron. Astrophys., 28, 174
Gopalswamy, N., Yashiro, S., Akiyama, S., et al. 2008, Annales Geophysicae, 26, 3033
Gopalswamy, N., Yashiro, S., Michałek, G., et al. 2002, ApJ, 572, L103
Jiang, L., Zhang, H., & Cai, Z. 2009, Knowledge and Data Engineering, IEEE Transactions on, 21, 1361
Jolliffe, I. T., & Stephenson, D. B. 2003, Forecast verification: a practitioner's guide in atmospheric science (Chichester: J. Wiley)
Kahler, S. W., Cliver, E. W., & Ling, A. G. 2007, Journal of Atmospheric and Solar-Terrestrial Physics, 69, 43
Kahler, S. W., & Vourlidas, A. 2005, Journal of Geophysical Research (Space Physics), 110, A12S01
Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. 1998, IEEE Trans. Pattern Anal. Mach. Intell., 20, 226
Kocharov, L., & Torsti, J. 2002, Sol. Phys., 207, 149
Kubo, Y., & Akioka, M. 2004, Space Weather, 2, S01002
Laurenza, M., Cliver, E. W., Hewitt, J., et al. 2009, Space Weather, 7, S04008

Lehtinen, N. J., Pohjolainen, S., Huttunen-Heikinmaa, K., et al. 2008, Sol. Phys., 247, 151
Marqué, C., Posner, A., & Klein, K.-L. 2006, ApJ, 642, 1222
Núñez, M. 2011, Space Weather, 9, S07003
Ohki, K. 2003, Sol. Phys., 213, 111
Park, J., Moon, Y.-J., Lee, D. H., & Youn, S. 2010, Journal of Geophysical Research (Space Physics), 115, A10105
Reames, D. V. 1999, Space Sci. Rev., 90, 413
Smart, D. F., & Shea, M. A. 1989, Advances in Space Research, 9, 281
Wang, J.-L. 2000, Chin. Astron. Astrophys., 24, 10
Yashiro, S., Akiyama, S., Gopalswamy, N., & Howard, R. A. 2006, ApJ, 650, L143
Yu, D., Huang, X., Hu, Q., et al. 2010, ApJ, 709, 321