Research in Astron. Astrophys. **2013** Vol. **13** No. **3**, 351–358
http://www.raa-journal.org *http://www.iop.org/journals/raa*

Research in Astronomy and Astrophysics

Solar flare prediction using highly stressed longitudinal magnetic field parameters *

Xin Huang and Hua-Ning Wang

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

Received 2011 September 19; accepted 2012 September 20

Abstract Three new longitudinal magnetic field parameters are extracted from *SOHO*/MDI magnetograms to characterize properties of the stressed magnetic field in active regions, and their flare productivities are calculated for 1055 active regions. We find that the proposed parameters can be used to distinguish flaring samples from non-flaring samples. Using the long-term accumulated MDI data, we build the solar flare prediction model by using a data mining method. Furthermore, the decision boundary, which is used to divide flaring from non-flaring samples, is determined by the decision tree algorithm. Finally, the performance of the prediction model is evaluated by 10-fold cross validation technology. We conclude that an efficient solar flare prediction model can be built by the proposed longitudinal magnetic field parameters with the data mining method.

Key words: Sun: magnetic fields — Sun: flares — methods: statistical

1 INTRODUCTION

The prediction of solar flares is an important unsolved problem in solar physics. Many parameters have been proposed to describe the non-potentiality and complexity of active regions which are considered as necessary conditions for solar eruptions. This led McIntosh (1990) to define the McIntosh classification system for sunspot groups. The sunspots are classified in 60 possible classifications using three components of the McIntosh system. However, these morphological parameters are simply proxies of the photospheric magnetic field, and the relationships between solar flares and properties of the magnetic field in active regions are more fundamental (Bornmann & Shaw 1994; Sammis et al. 2000). Therefore, many magnetic field parameters that describe active regions are proposed to distinguish flaring samples from non-flaring samples. These parameters include the gradient of the magnetic fields (Krall et al. 1982; Leka & Barnes 2003), the length of the neutral line (Falconer et al. 2002; Cui et al. 2006), the number of singular points (Cui et al. 2006), the current density (Zhang 2001), the current helicity (Bao et al. 1999), the shear angle (Hagyard et al. 1984) and the magnetic twist (Zhang et al. 2002; Hahn et al. 2005). Leka & Barnes (2007) pointed out that none of these variables by themselves make a strong distinction between flaring samples and non-flaring samples, therefore many efforts were continually made to improve the performance of solar flare prediction. Barnes & Leka (2006) quantified the topological complexity describing the coronal fields associated with an active region. Georgoulis & Rust (2007) defined the effective connected magnetic field

* Supported by the National Natural Science Foundation of China.

to measure the flaring potential in active regions. Welsch et al. (2009) quantitatively characterized the flow field of active regions. Yu et al. (2009) studied influences of parameter sequences on the prediction of solar flares. Huang et al. (2010) proposed subsets of predictors to group complementary parameters together. Komm et al. (2011) proposed subsurface flow parameters to distinguish between flaring and non-flaring samples.

Large-sample studies based on long-duration observations of active regions are important for the prediction of solar flares (Schrijver 2009), and the accumulated data from the Michelson Doppler Imager (MDI) onboard the *Solar and Heliospheric Observatory* (*SOHO*) are suitable for this purpose. However, most of the above-mentioned works focus on properties of the photospheric vector magnetic field with a relatively small number of samples. Hence, we propose three new photospheric longitudinal magnetic field parameters from MDI magnetograms to describe the stressed property of active regions. The relationships between the proposed parameters and solar flares were studied by a large-sample statistical method in which more than 70 000 samples of 1055 active regions are observed. Using this large-sample dataset, a short-term solar flare prediction model is built using a data mining method, and its performance is evaluated by 10-fold cross validation technology.

This paper is organized as follows: the data are introduced in Section 2, the prediction model is built and evaluated in Section 3, and conclusions are given in Section 4.

2 DATA

In order to build a solar flare prediction model by using a data mining method (Wang et al. 2008; Qahwaji et al. 2008; Li et al. 2008; Yu et al. 2010; Ball & Brunner 2010; Huang et al. 2012a; Huang et al. 2012b), a large dataset is required.

2.1 Extraction of an Active Region and Definition of a Flare Index

Mason & Hoeksema (2010) provided the location and the extension of active regions within 30° of the solar disk's center from 1996 May 10 to 2007 June 9[1]. Hence, data describing 1055 active regions in 70 078 magnetograms (Scherrer et al. 1995) are extracted to form the dataset.

According to the peak flux of soft X-rays, solar flares are classified as C, M or X. For each classification, a linear scale (1 to 9) is used to determine the exact value of the solar flare. Within a forward looking period, more than one flare may occur, and the total importance (Antalova 1996) of these flares is defined as

$$I_{\rm tot} = \sum c + 10 \times \sum m + 100 \times \sum x, \tag{1}$$

where c, m and x represent linear scales in terms of solar flare classifications C, M and X, respectively. If the total importance of solar flares within a forward looking period is larger than a given threshold, this sample observation of the active region is considered to be flaring. Otherwise, it is considered to be non-flaring. Here, the threshold of $I_{\rm tot}$ is set to be 10 (M1.0 equivalent), and the forward looking period is taken to be 48 hours.

2.2 Highly Stressed Longitudinal Magnetic Field Parameters

Wang (1995) proposed a parameter P to quantify the highly stressed longitudinal magnetic fields in active regions.

$$P = \nabla_{\rm h} B_z \cdot \frac{\boldsymbol{B}_{\rm pt}}{|\boldsymbol{B}_{\rm pt}|}, \tag{2}$$

[1] *http://soi.stanford.edu/data/tables/*

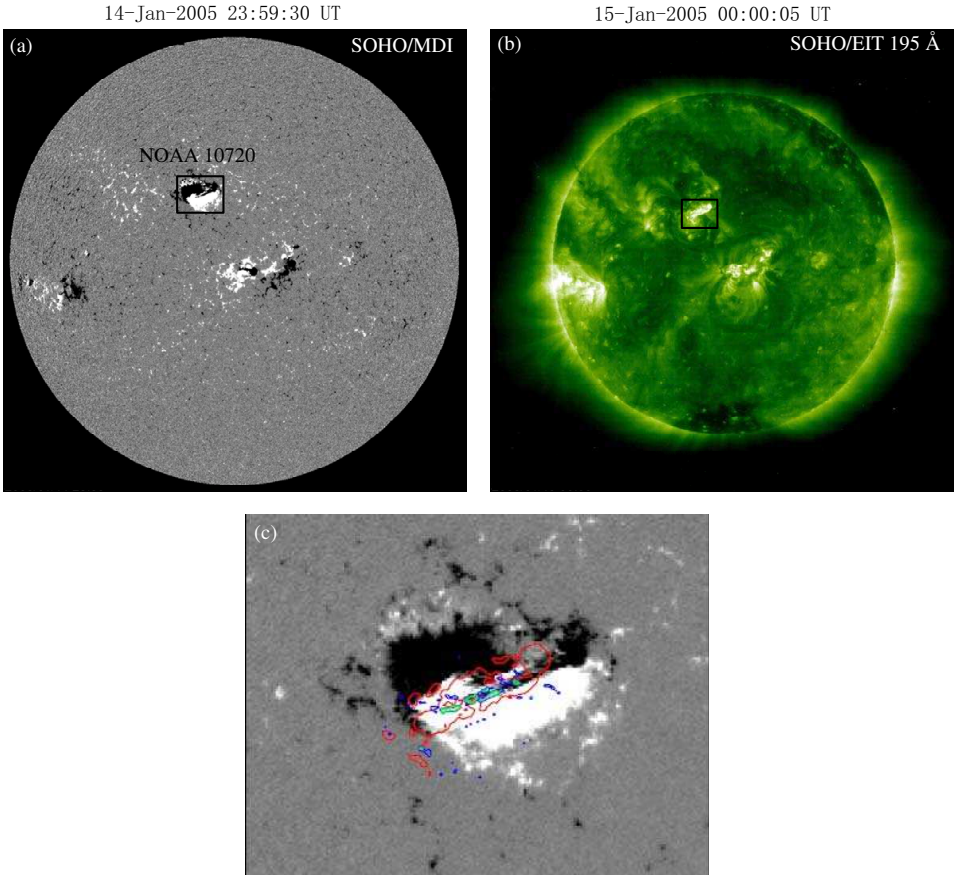

Fig. 1 (a) Full disk image of *SOHO*/MDI magnetogram observed on 2005 January 14 at 23:59:30 UT, and the active region NOAA 10720 is marked by the black rectangle. (b) Full disk *SOHO*/EIT 195 Å image observed on 2005 January 15 at 00:00:05 UT. (c) Isocontours of the solar flare and the stressed longitudinal magnetic field parameter P overplotted on the *SOHO*/MDI magnetogram for the active region NOAA 10720. The background is the grayscale map of the longitudinal magnetic field. Red contours enclose the solar flare observed by the *SOHO*/EIT 195 Å image. Blue and green contours show 0.1 G km^{-1} and 0.2 G km^{-1} for the value of P, respectively.

where $\nabla_{\rm h} B_z$ is the horizontal gradient of the longitudinal magnetic field, and $\boldsymbol{B}_{\rm pt}$ is the transverse component of the magnetic field inferred by the potential field model with the boundary condition of the longitudinal magnetic field.

The direction of $\nabla_{\rm h} B_z$ is usually different from that of $\boldsymbol{B}_{\rm pt}$, however it changes when the longitudinal magnetic field is stressed. Hence, positive P indicates the presence of highly stressed longitudinal magnetic fields. Taking the active region NOAA 10720 (the locations of this region in the *SOHO*/MDI magnetogram and *SOHO*/EIT 195 Å image are shown in Figure 1(a) and (b), respectively) as an example, we show isocontours of the solar flare and the stressed longitudinal magnetic field parameter P overplotted on the *SOHO*/MDI magnetogram in Figure 1(c). We find that the solar flare appears in the place where P is large. Therefore, we propose three parameters to characterize the high stress of the longitudinal magnetic field in an active region:

(1) $P_{\rm num}$: the number of pixels where the P is positive in an active region.
(2) $P_{\rm max}$: the maximum of P in an active region.
(3) $P_{\rm sum}$: the summation of P where P is positive in an active region.

The flare productivity (FP) for a parameter (x) is defined as

$$\mathrm{FP}(X) = \frac{N_{\rm Flare}(x \geq X)}{N_{\rm Total}(x \geq X)}, \tag{3}$$

where X is a given threshold for a magnetic field parameter x, $N_{\rm Flare}(x \geq X)$ and $N_{\rm Total}(x \geq X)$ are the number of flaring samples and the total number of samples when the value of parameter x is larger than the threshold X, respectively.

The threshold X is determined by the number of eliminated samples whose value is smaller than the given threshold rather than equally dividing the range of the parameter. Taking the parameter $P_{\rm sum}$ for example, the initial flare productivity calculated by all the samples is 15.15%, which means that the threshold X is at its minimum. We increase the threshold X until 500 samples are included and the value of $P_{\rm sum}$ becomes smaller than the threshold X. The corresponding threshold X of $P_{\rm sum}$ is 0.18 (G km^{-1}) and the flare productivity is 15.17%. This process is repeated until the remaining samples are less than 500. The generated flare productivity curves for the proposed parameters are shown in Figure 2. We find that by increasing the proposed parameters, the corresponding flare productivity increases, so these parameters can be used for distinguishing flaring samples from non-flaring samples.

3 SOLAR FLARE PREDICTION MODEL

3.1 Building a Model from Data

There are two methods which can be used to build the prediction model:

(1) The deductive method: it starts with physical laws to provide relationships between magnetic parameters and solar flares.
(2) The inductive method: it starts with observations of physical phenomena to summarize relationships between magnetic parameters and solar flares.

The decision tree algorithm (Quinlan 1996; Yu et al. 2009) is one inductive method, and the divide-and-conquer strategy is used to construct the decision tree model from the dataset. In this strategy, a best parameter is selected in each step to divide samples into smaller subsets. We hope that samples with the same classification are assigned to the same subset as much as possible, so the importance of a parameter is measured by the information gain ratio (GR)

$$\mathrm{GR}(y, X_i) = \frac{H(y) - H(y|X_i)}{H(X_i)}, \tag{4}$$

where y is the status of solar flares, X_i stands for the ith parameter ($i = 1, 2, 3$ stands for $P_{\rm num}$, $P_{\rm max}$ and $P_{\rm sum}$ respectively.), and H stands for the information entropy (Yu et al. 2009) which describes the uncertainty of the system. The information GR measures the reduction in the uncertainty of samples associated with the partition process.

For a continuous parameter (CP), samples are sorted by their values to give the ordered distinct values dv_1, dv_2, ..., dv_N (for example, 5.7, 5.9, 6.0 etc. shown in Table 1). Every pair of adjacent values suggests a possible threshold $\mathrm{PT} = \frac{(dv_i + dv_{i+1})}{2}$. The threshold T, which is used to divide the dataset into two parts (CP$\leq$ T and CP$>$ T), should be determined by maximizing the information GR among the possible thresholds (Quinlan 1996). Suppose we use $P_{\rm sum}$ to distinguish flaring samples from non-flaring samples (see Table 1). We sort samples by the value of $P_{\rm sum}$, and the possible

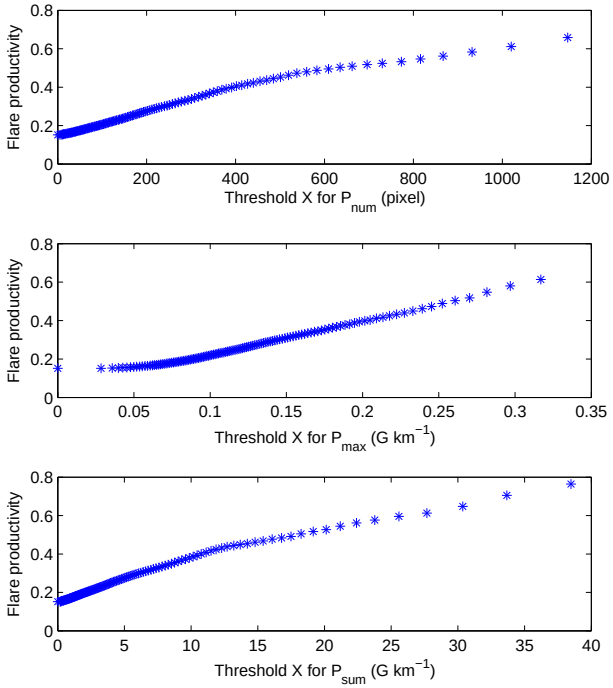

Fig. 2 Flare productivity curves of $P_{\rm num}$, $P_{\rm max}$ and $P_{\rm sum}$ for different threshold values X. The first point is calculated by all the samples. With the increase of the threshold X, the number of remaining samples decreases. The threshold X is determined by the number of eliminated samples whose value is smaller than the given threshold. At each step 500 samples are reduced, and this process is repeated until the number of remaining samples is less than 500.

Table 1 Example of Possible Positions to Split the Samples

$P_{\rm sum}$ (G km^{-1})	5.7	5.9	6.0	6.1	6.2	6.2	6.4
Solar flares	0	1	0	0	1	1	1

Note: There are three possible positions ($\frac{5.7+5.9}{2}$,$\frac{5.9+6.0}{2}$ and $\frac{6.1+6.2}{2}$) to slit the samples. The partition value with the largest information GR is selected as the final partition threshold. "1" and "0" stand for flaring and non-flaring samples, respectively.

positions to split samples are determined by changes in the status for solar flares. There are three possible positions ($\frac{5.7+5.9}{2}$, $\frac{5.9+6.0}{2}$ and $\frac{6.1+6.2}{2}$) to split the samples in Table 1. The information GR for each possible partition value is calculated and the position with the largest value of information GR is used to divide the samples into the subsets. The best parameter determined by the information GR is selected to divide the samples into the smaller subsets, and this process is recursive until the stopping criterion is satisfied. The stopping criterion is that the subset contains only one type of sample (flaring samples or non-flaring samples) or there are no improvements of information GR for the next possible partition. At this point, samples in the subset are considered to represent the class with the most samples in this subset.

The decision boundary generated by the decision tree algorithm for solar flare prediction using the parameters $P_{\rm num}$, $P_{\rm max}$, and $P_{\rm sum}$ is shown in Figure 3.

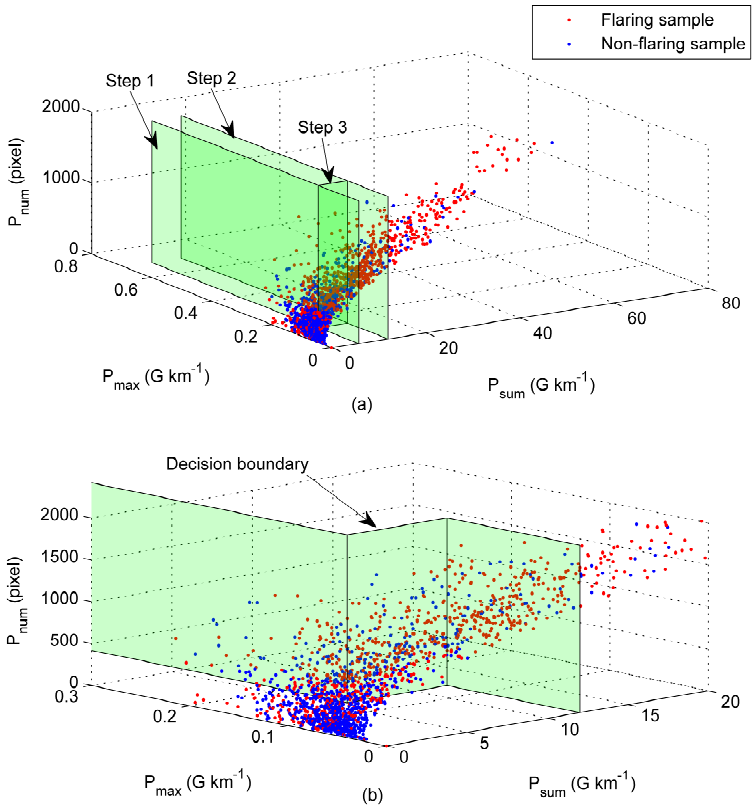

Fig. 3 Decision boundary generated by the decision tree algorithm in parameter space of $P_{\rm num}$, $P_{\rm max}$ and $P_{\rm sum}$. The decision tree algorithm divides the samples into subsets indicated by the parameter with the largest information GR at each step, and this process is repeated until the stopping criterion is achieved. (a) shows the generation process of the decision boundary. The final decision boundary is shown in (b), and the left side is considered to be a non-flaring area, while the right side is considered to be a flaring area.

Figure 3(a) shows the associated generation process in detail. In step 1, $P_{\rm sum}$ with the partition threshold T$=6$ G km^{-1} is selected to divide samples into two subsets. The samples whose $P_{\rm sum}$ is smaller than 6 G km^{-1} satisfy the stopping criterion, so these samples do not need further division. The samples whose $P_{\rm sum}$ is larger than 6 G km^{-1} continue to be divided by $P_{\rm sum}$ with the partition threshold T$=12$ G km^{-1} in step 2 and $P_{\rm max}$ with the partition threshold T$=0.1$ G km^{-1} in step 3, until all the subsets satisfy the stopping criterion. Finally, the generated decision boundary is shown in Figure 3(b), in which samples on the left side of the decision boundary are considered to be non-flaring samples, and samples on the right side of the decision boundary are considered to be flaring samples.

The whole process can be represented by a tree-like structure, so the prediction model is called a decision tree model. In Figure 3, we find that $P_{\rm num}$ is not used in the prediction model, because given the selected parameters, $P_{\rm num}$ does not contain more information for distinguishing between flaring samples and non-flaring samples. From the prediction model, we find that the active region with a more highly stressed longitudinal magnetic field is more prone to produce solar flares.

Table 2 Definition of Contingency Table for Solar Flare Prediction

	Predicted flaring samples	Predicted non-flaring samples
Actual flaring samples	Hit	Miss
Actual non-flaring samples	False alarm	Correct rejection

Table 3 Testing Results of the Solar Flare Prediction Model in the Contingency Table

	Predicted flaring samples	Predicted non-flaring samples
Actual flaring samples	686 ± 26	254 ± 26
Actual non-flaring samples	1609 ± 153	3941 ± 153

3.2 Performance Evaluation

The performance of the prediction model can be estimated by a 10-fold cross validation technique. In this technique, the dataset is randomly divided into 10 folds with approximately equal size; 9 folds are considered as the training set, and the remaining fold is used as the testing set. The prediction model is built from the training set and its performance is estimated on the testing set. This process is repeated 10 times, until each of the 10 subsets is used exactly once as the testing set. Finally, the 10 results are averaged to produce a single estimation.

The solar flare prediction model has four possible outputs, and its contingency table (Jolliffe & Stephenson 2003) is defined in Table 2. Based on the contingency table, the hit rate and the correct rejection rate are defined to measure the performance of flaring prediction and non-flaring prediction, respectively.

$$\text{Hit rate} = \frac{N_{\rm hit}}{N_{\rm hit} + N_{\rm miss}}, \tag{5}$$

where $N_{\rm hit}$ is the number of hit samples and $N_{\rm miss}$ is the number of miss samples.

$$\text{Correct rejection rate} = \frac{N_{\rm CR}}{N_{\rm CR} + N_{\rm FA}}, \tag{6}$$

where $N_{\rm CR}$ is the number of correct rejection samples and $N_{\rm FA}$ is the number of false alarm samples.

The results from testing the solar flare prediction model are shown in Table 3; furthermore, the hit rate of the prediction model is $73\% \pm 3\%$ and the correct rejection rate of the prediction model is $71\% \pm 3\%$. Although the solar flare prediction model based on the proposed three parameters cannot completely distinguish flaring samples from non-flaring samples, the performance of this model is similar to that of an experienced forecaster (Wang et al. 2008).

4 CONCLUSIONS

Extracting three parameters ($P_{\rm num}$, $P_{\rm max}$ and $P_{\rm sum}$) from *SOHO*/MDI magnetograms, we quantify the stress of a longitudinal magnetic field in active regions, and calculate the flare productivity for each parameter. We find that the proposed parameters effectively characterize the stress of the longitudinal magnetic field and these parameters can be used to distinguish flaring samples from non-flaring samples.

In the parameter space of $P_{\rm num}$, $P_{\rm max}$ and $P_{\rm sum}$, the decision boundary, which is used to divide the flaring samples from non-flaring samples, is generated by the decision tree algorithm. The decision tree algorithm automatically determines the locally optimal partition threshold at each step. The decision tree model, which represents this partition process, may be helpful for understanding the physical basis of the solar flare prediction. The hit rate and the correct rejection rate of the prediction model are both larger than 70%, so the derived model is efficient for prediction of solar flares.

Generally speaking, the proposed parameters characterize the non-potentiality of active regions. In the future, the trigger mechanism of solar flares should be studied to improve the performance of the solar flare prediction model.

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 National Basic Research Program of China (973 Program, Grant No. 2011CB811406), the National Natural Science Foundation of China (Grant Nos. 11273031, 10733020, 10921303 and 11078010) and the China Meteorological Administration grant (No. GYHY201106011). This paper has benefited from comments given by the reviewer.

References

Antalova, A. 1996, Contributions of the Astronomical Observatory Skalnate Pleso, 26, 98
Ball, N. M., & Brunner, R. J. 2010, International Journal of Modern Physics D, 19, 1049
Bao, S. D., Zhang, H. Q., Ai, G. X., & Zhang, M. 1999, A&AS, 139, 311
Barnes, G., & Leka, K. D. 2006, ApJ, 646, 1303
Bornmann, P. L., & Shaw, D. 1994, Sol. Phys., 150, 127
Cui, Y., Li, R., Zhang, L., He, Y., & Wang, H. 2006, Sol. Phys., 237, 45
Falconer, D. A., Moore, R. L., & Gary, G. A. 2002, ApJ, 569, 1016
Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 661, L109
Hagyard, M. J., Smith, J. B., Teuber, D., & West, E. A. 1984, Sol. Phys., 91, 115
Hahn, M., Gaard, S., Jibben, P., Canfield, R. C., & Nandy, D. 2005, ApJ, 629, 1135
Huang, X., Yu, D., Hu, Q., Wang, H., & Cui, Y. 2010, Sol. Phys., 263, 175
Huang, X., Wang, H., & Dai, X. 2012a, Science in China: Phys Mech Astron, 55, 1
Huang, X., Wang, H.-N., & Li, L.-P. 2012b, RAA (Research in Astronomy and Astrophysics), 12, 313
Jolliffe, I., & Stephenson, D. 2003, Forecast Verification: A Practitioner's Guide in Atmospheric Science (Wiley Online Library)
Komm, R., Ferguson, R., Hill, F., Barnes, G., & Leka, K. D. 2011, Sol. Phys., 268, 389
Krall, K. R., Smith, J. B., Jr., Hagyard, M. J., West, E. A., & Cummings, N. P. 1982, Sol. Phys., 79, 59
Leka, K. D., & Barnes, G. 2003, ApJ, 595, 1277
Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173
Li, R., Cui, Y., He, H., & Wang, H. 2008, Advances in Space Research, 42, 1469
Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634
McIntosh, P. S. 1990, Sol. Phys., 125, 251
Qahwaji, R., Colak, T., Al-Omari, M., & Ipson, S. 2008, Sol. Phys., 248, 471
Quinlan, J. R. 1996, Journal of Artificial Intelligence Research, 4, 77 (arXiv:cs/9603103)
Sammis, I., Tang, F., & Zirin, H. 2000, ApJ, 540, 583
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, Sol. Phys., 162, 129
Schrijver, C. J. 2009, Advances in Space Research, 43, 739
Wang, H. 1995, Sol. Phys., 157, 213
Wang, H. N., Cui, Y. M., Li, R., Zhang, L. Y., & Han, H. 2008, Advances in Space Research, 42, 1464
Welsch, B. T., Li, Y., Schuck, P. W., & Fisher, G. H. 2009, ApJ, 705, 821
Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, Sol. Phys., 255, 91
Yu, D., Huang, X., Wang, H., et al. 2010, ApJ, 710, 869
Zhang, H. 2001, ApJ, 557, L71
Zhang, H., Bao, S., & Kuzanyan, K. M. 2002, Astronomy Reports, 46, 424