MAG4 versus Alternative Techniques for Forecasting Active-Region Flare Productivity

David A. Falconer,[1,3] Ronald L. Moore,[1,3] Abdulnasser F. Barghouty,[1,2] and Igor Khazanov[1,3]

1. Heliophysics and Planetary Science Office, Space Weather Team, ZP13 MSFC/NASA, Huntsville Al, 35812
2. Astrophysics Office ZP12 MSFC/NASA,
3. Center for Space Plasma and Aeronomic Research, The University of Alabama in Huntsville
4. Corresponding Author: David A. Falconer Heliophysics and Planetary Office ZP13 MSFC/NASA, Huntsville Al, 35812 David.a.falconer@nasa.gov

Key Points

- Quantitative comparison of performance of pairs of forecasting techniques
- Next MAG4 forecasts major flares more accurately than Present MAG4
- Present MAG4 forecasts more accurately than either McIntosh AR Class or Total Magnetic Flux

Abstract

MAG4 is a technique of forecasting an active region's rate of production of major flares in the coming few days from a free-magnetic-energy proxy. We present a statistical method of measuring the difference in performance between MAG4 and comparable alternative techniques that forecast an active region's major-flare productivity from alternative observed aspects of the active region. We demonstrate the method by measuring the difference in performance between the "Present MAG4" technique and each of three alternative techniques, called "McIntosh Active-Region Class," "Total Magnetic Flux," and "Next MAG4." We do this by using (1) the MAG4 database of magnetograms and major-flare histories of sunspot active regions, (2) the NOAA table of the major-flare productivity of each of 60 McIntosh active-region classes of sunspot active regions, and (3) five technique-performance metrics (Heidke Skill Score, True Skill Score, Percent Correct, Probability of Detection, and False Alarm Rate) evaluated from 2000 random two-by-two contingency tables obtained from the databases. We find that (1) Present MAG4 far outperforms both McIntosh Active-Region Class and Total Magnetic Flux, (2) Next MAG4 significantly outperforms Present MAG4, (3) the performance of Next MAG4 is insensitive to the forward and backward temporal windows used, in the range of one to a few days, and (4) forecasting from the free-energy proxy in combination with either any broad category of McIntosh active-region classes or any Mount Wilson active-region class gives no significant performance improvement over forecasting from the free-energy proxy alone (Present MAG4).

Index terms: Space Weather Forecasting, Solar Flares, Solar Magnetic fields

1. Introduction

Solar flares and coronal mass ejections (CMEs) are the primary drivers of severe space weather. Devising better techniques of forecasting flares and CMEs is essential for better forecasting of space weather, severe or mild. To determine whether a newly-devised technique is significantly better than an alternative similar technique, it is crucial to have a way of measuring both the difference in performance of the two techniques and the statistical significance of the difference. This paper demonstrates a method of measuring the difference in performance of different forecasting techniques of the same type. We apply this method to measuring the difference in performance of many pairs of techniques of forecasting X- and M-class flares. The method could also be used to compare the performance of these same pairs of techniques when these are applied to the forecasting of CMEs. Because our database of X- and M-class flares that have known source active regions (ARs) is larger than our database of CMEs that have known source ARs, in this paper we limit our demonstration of our method of measuring and comparing the performance of similar techniques to the performance of the techniques in forecasting X- and M- flares.

For forecasting X- and M-class flares, NOAA presently uses the McIntosh active-region (AR) classes (McIntosh 1990) as the basis of their forecasts. Each of the 60 different McIntosh AR classes has a different three-letter designation reflecting the following attributes of the AR: unipolar/bipolar, penumbra presence, longitudinal extent, penumbra development, penumbra asymmetry, leading spot diameter, and how filled-in the active region is with sunspots (see the first two columns of Table 5 in Appendix B for a list of attributes). NOAA uses a database spanning 1996-2006 to assign to any active region of a given McIntosh class a probability or event rate for having X-class, M-class, or C-class flares.

Besides using the McIntosh AR classes for forecasting, there have been several other attempts at producing better flare forecasts using aspects of active regions. These include previous flare activity (Wheatland 2004, 2005), magnetic complexity (Abramenko 2005; Georgoulis 2012; McAtteer et al. 2005), helioseismology signatures (Komm et al. 2005; Reinard et al. 2010), photospheric flows (Welsch et al. 2009), total magnetic flux (Barnes & Leka 2008; Leka & Barnes 2007), free-energy proxies (Falconer 2001; Falconer et al. 2002, 2003, 2006, 2008, 2011; Leka & Barnes 2003a,b; Cui et al. 2006; Jing et al. 2006; Georgoulis & Rust 2007; Schrijver 2007; Mason & Hoeksema 2010), and the combination of a free-energy proxy and previous flare history (Falconer et al. 2012).

MAG4 was developed to assist NASA/SRAG (Space Radiation Analysis Group at Johnson Space Flight Center) in forecasting X- and M-class flares, CMEs, and Solar Proton Event (SPE). It forecasts an AR's rate of producing each kind of event from a proxy of the AR's free magnetic energy, which proxy is measured from a magnetogram of the AR. This free-energy proxy is developed and defined in a series of papers (Falconer et al. 2002, 2003, 2006, 2008, 2009, 2011, 2012). Falconer et al. (2011) describes in detail how MAG4 works. Falconer et al. (2012) shows that using the AR's previous flare activity together with its free-energy proxy gives a better forecast than that obtained from the free-energy proxy alone. In this paper, we will quantitatively compare the performance of four forecasting techniques that are based on different aspects of active regions: (1) the McIntosh AR class (the McIntosh AR Class technique), (2) the AR's total magnetic flux (the Total Magnetic Flux technique), (3) the AR's free-energy proxy only (the Present MAG4 technique), and (4) the free-energy proxy combined with the AR's previous flare activity (the Next MAG4 technique). Present MAG4 uses empirical forecasting curves to convert an AR's measured free-energy proxy into the AR's forecasted rate of

production of each kind of event (Falconer et al. 2011). The forecasting curves are derived from a sample of 40,000 magnetograms of 1,300 ARs observed by the SOHO/Michelson Doppler Imager (MDI) line-of-sight magnetograph (Scherrer et al. 1995) between 1996 and 2004 and from the observed histories of production of each kind of event by these active regions.

To measure the difference in performance of the above four forecasting techniques, we do the following. We 1) for each of 2000 runs, randomly divide the 40,000 magnetogram sample into two halves, a control subsample and an experimental subsample, assigning all magnetograms of the same active region to one or the other subsample (so that the two subsamples are independent); 2) for each run, derive for each of three of the forecasting techniques the forecasting curve for X- and M-class flares from the control subsample; 3) apply this forecasting curve, or in case of the technique based on the McIntosh AR classes, apply the NOAA lookup tables, to the experimental subsample; 4) for each of the four techniques, from each run obtain a 2x2 contingency table by converting each AR magnetogram's forecasted AR event rate to a forecast of "yes" or "no" for expected production of an X- or M- flare by the AR during a forward window of one or more days from the time of the magnetogram; 5) evaluate five different performance metrics from the 2x2 contingency table; 6) compare the average and standard deviation of each performance metric for pairs of forecasting techniques; 7) compare the average and standard deviation of the difference of each metric for pairs of forecasting techniques.

We find that McIntosh AR Class and Total Magnetic Flux have similar performance metrics. We also find that the Present MAG4 technique is statistically superior in performance to both the McIntosh AR Class technique and the Total Magnetic Flux technique. In addition, Next MAG4 is statistically superior in performance to Present MAG4. We also show that there is no choice of duration of the backward temporal window to categorize an active region as recently flaring or not that significantly improves the performance of Next MAG4 over that for a 24 hr window, for choices in the range of 24-96 hours, showing that the result for the choice of 24 hours is robust. Similarly, the performance of Next MAG4 is insensitive to the duration of the forward window used to determine the event-rate forecasting curves (Appendix A). We find no forecasting technique that is based on any combination of the free-energy proxy and broad categories of the McIntosh AR classes gives any significant improvement in the performance compared to Present MAG4, and that these techniques often result in poorer performance (Appendix B).

The organization of the remainder of this paper is as follows: In Section 2, we describe the magnetic measures used in this study. In Section 3, we describe how we convert these magnetic measures to forecasted event rates via forecasting curves. In Section 4, we describe how we obtain 2x2 contingency tables from which we evaluate five performance metrics for each forecasting technique, and then compare the average and standard deviation of these metrics from 2000 Monte-Carlo runs for each of the four forecasting techniques. In Section 5, we compare the difference in forecasting performance between pairs of forecasting techniques shown by their same-run differences in each performance metric. In Section 6, we present our conclusions. In Appendix A, we search for the best temporal windows to use for the Next MAG4 technique. In Appendix B, we explore whether any significant improvement in forecasting performance can be obtained by combining free-energy proxy with some McIntosh or Mount Wilson attribute of active regions.

2. Magnetogram Measures

The AR magnetogram database used in this paper is from the MDI magnetograms from 1996-2004, those used by the MAG4 (Magnetogram Forecast) forecasting tool

(http://www.uah.edu/cspar/research/mag4-page), which is described in Falconer et al. (2011). MAG4 downloads an SDO/HMI (Solar Dynamics Observatory/Helioseismic and Magnetic Imager), GONG (Global Oscillation Network Group), or MDI full-disk magnetogram, isolates the strong magnetic field areas in it, and matches each of these areas to one or more NOAA active regions, if any. MAG4 then evaluates the free-energy proxy (described below) for each AR, and from that makes the forecast for each AR (described in Section 3). The 1996-through-2004 period was chosen for the database because we had the flare and CME histories of the active regions of that period (tables from NOAA, Chris Balch, personel communication 2007), which were double checked by us (Falconer et al. 2011). We chose only active-region magnetograms observed within 30 heliocentric degrees of disk center to keep projection effects small. In Falconer et al. (2008) we determined from vector magnetograms that our free-energy proxy measured from the line-of-sight component of the vector magnetogram begins having unacceptable projection errors for AR distances beyond 30 heliocentric degrees from disk center. All magnetic measurements used in this paper are from the line-of-sight magnetograms from MDI, and are evaluated using the line-of-sight (LOS) approximation (treating the line-of-sight component of the magnetic field as if it were the vertical component of the magnetic field). We also used only those strong-field areas that were assigned to only one NOAA active region, i.e., any strong-field area containing two or more NOAA ARs was not used. The resulting sample is about 40,000 magnetograms, from about 1,300 separate NOAA active regions, with each active region sampled up to once every 96 minutes, the MDI cadence; with observations of each active region spanning up to 5 days for low-latitude active regions (see Falconer et al. (2011) for more details).

There are four magnetogram measures used in this paper. These are the AR's gradient-weighted length of strong-field neutral line, the AR's length of strong-field neutral line, the AR's total magnetic flux, and the AR's magnetic area. The latter two measures are from all pixels of the AR magnetogram that have absolute LOS magnetic field of greater than 100 G. The magnetic area is then,

$$^{L}A_{m} = \int da, \tag{1}$$

where the integral is over all pixels with absolute magnetic field greater than 100 G. The total magnetic flux is defined as

$$^{L}\Phi = \int |B_{LOS}| da \tag{2}$$

where $|B_{LOS}|$ is the magnitude of the line-of-sight magnetic field and again the integral is over all pixels with LOS field strength greater than 100 G. The superscript L indicates that we are using the line-of-sight approximation.

The neutral line (or polarity inversion line) in an AR magnetogram is the line that separates positive and negative polarity flux, i.e., the zero-Gauss contour. We divide the zero-Gauss contour into segments, each roughly a pixel in length. For each segment, its length, neighbor segments, the interpolated values of the potential transverse field at the midpoint of the segment, the transverse gradient of the line-of-sight field at the midpoint, the average positive field nearby, and the average negative field nearby are compiled. Those segments are kept that meet the following two conditions: 1) it has strong potential transverse field (> 150 G); 2) it separates at least moderately strong positive and negative field (five-pixel-smoothed flux of either polarity is stronger than 15 G). The potential transverse field is calculated from the observed line-of-sight magnetogram of the AR (Alissandrakis, 1981).) All kept segments are checked for whether it has a neighbor kept segment; isolated kept segments are then eliminated.

The rules for average positive and negative field strength ensure that the segment separates active-region polarities instead of an active-region polarity and a quiet-sun polarity.

The other two magnetogram measures are neutral-line measures. One is the length of strong-field neutral line,

$$^{L}L_S = \int dl, \qquad (3)$$

where the integral is over those neutral-line segments that fulfill the above three conditions (listed in the previous paragraph) in the AR's line-of-sight magnetogram (see the example AR magnetogram in Figure 1), comprising the strong-field intervals of the neutral lines. The other neutral-line measure is a proxy of the AR's free magnetic energy:

$$^{L}WL_{SG} = \int |\nabla B_{LOS}| dl \qquad (4)$$

where $^{L}WL_{SG}$ is the gradient-weighted length of strong-field neutral line (the LOS approximation).

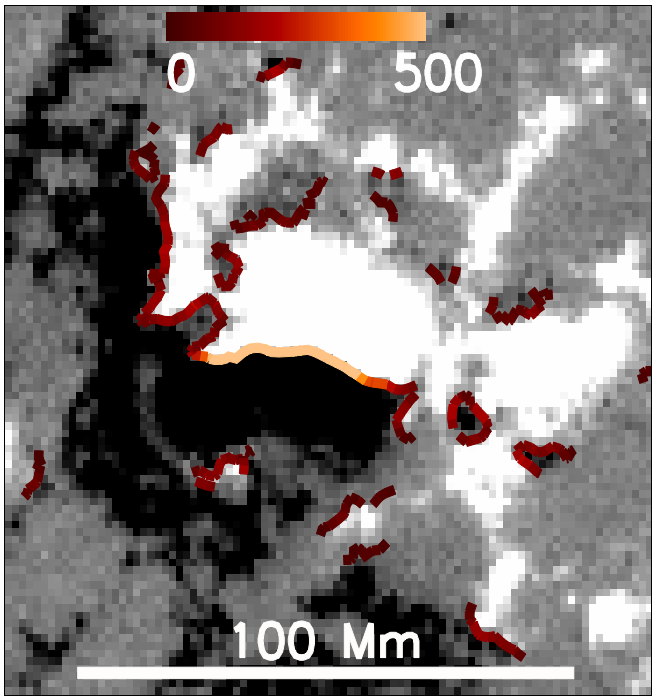

Figure 1 The strong-field intervals of the neutral lines in a MDI magnetogram of a δ-sunspot active region that produced an X-class flare and CME. The polarity, strength, and distribution of the line-of-sight flux are mapped by the grayscale image. The polarity is positive (negative) in light (dark) areas. The strong-field intervals of the neutral lines are traced by the colored curves. The color gives the strength of the gradient of the line-of-sight field ($|\nabla B_{LOS}|$) at these neutral lines in units of G/Mm. The lightest color is for $|\nabla B_{LOS}| \geq 500$ G/Mm, the range of the extreme gradients at the interval of the neutral line in the large δ sunspot.

In previous work we have shown that $^{L}WL_{SG}$ is a proxy of the nonpotentiality of the magnetic field or the free-magnetic energy. In Falconer et al (2003) we have shown that near disk center $^{L}WL_{SG}$ is a good approximation of the more physical WL_{SG} ($= \int |\nabla B_Z| dl$), and is strongly correlated with the corresponding neutral-line integral of the magnetic shear angle (the angle between the potential and observed transverse field). In Falconer et al (2006) we further showed the correlation of WL_{SG} with the net electric current that flows from one polarity to the other in bipolar active regions.

Only strong-field active regions are kept for the database. We define a strong-field active region to be any active region for which the ratio of the length of strong-field neutral line to the square root of the magnetic area $[^{L}L_S/(^{L}A_m)^{0.5}]$ is greater than 0.75. Keeping only the strong-field active regions eliminates the old decaying active regions from which we cannot make adequate measurements of the free-energy proxy.

3. The Four Forecasting Techniques

We will measure the performance of four comparable forecasting techniques. These four alternative techniques differ in three ways: 1) one technique forecasts from discrete categories of ARs; 2) two of the techniques forecast from a single parameter, either AR total magnetic flux or AR gradient-weighted length of strong-field neutral line ($^{L}WL_{SG}$); 3) the fourth technique forecasts from $^{L}WL_{SG}$ and a binary discriminator, namely whether or not the AR recently

produced a major flare. The first forecasting technique is based on the McIntosh AR classes (McIntosh 1990, Bornmann, Kalmbach, and Kulhanek 1994, Bornmann and Shaw 1994, and Appendix B). Here each active region is assigned to one of 60 classes based on whether the active region is unipolar or bipolar, whether or not there are penumbras, whether there is penumbra on sunspots at one end or both ends of the AR, length of the AR, maturity of penumbra, symmetric/asymmetric penumbra, diameter of the largest spot, how much the AR is filled in with sunspots between the leading and trailing spot, and whether an interior sunspot has a mature penumbra. Several classes are quite rare, having less than 10 active-region days between 1996 and 2006. For each class NOAA has compiled the average number of X-class, M-class or greater, and C-class or greater flares that the ARs of that class produced from August 1996 to March 2006. This database yields the average flare rate for each AR class (Balch, private communication: this tabulation is here after referred to as the NOAA Table), and this is the flare rate forecast by the McIntosh AR Class technique for each new AR of that class. In practice at NOAA, the flare rate forecast from McIntosh AR class can be adjusted by the forecaster, based on the forecaster's years of experience of forecasting from AR morphology. Here we are measuring the performance of the forecasting technique based on the McIntosh AR classes (McIntosh AR Class) with no adjustments to the forecast flare rates made by a forecaster.

To make forecasts based on a single parameter (such as the AR's total magnetic flux or the AR's free-energy proxy), we sort the active regions in our database by that parameter into up to 40 equally populated bins (see Falconer et al. 2011). For each bin, we calculate the average event rate and average value of the parameter. The flare rate is obtained from the database by using a forward temporal window of 24 hours (T_f = 24 hr) from the time of each AR magnetogram. From the number of events we obtain the likely parent-population event rate assuming Poisson statistics (Sachs 1978). We fit these bin rates assuming a functional form,

$$R = A\, X^B \tag{5}$$

where X is the parameter, "*A*" and "*B*" are empirically determined fitting parameters, and R is the forecasted X- and M-class flare rate. Because most bins having ${}^{L}WL_{SG} < 10^4$ G or ${}^{L}\Phi < 10^{22}$ Mx have no X- or M-class flare in them, we fit only bins having values larger than these.

To make a forecast based on two parameters (as in Next MAG4), where the second parameter has only two discrete states, we first group the sample by the second parameter: for example, whether the active region has or has not produced an X- or M-class flare in the previous 24 hours (T_b = 24 hr) before the time of the magnetogram, and for each group obtain a separate forecasting curve (Figure 2). That is, in this example, we obtain one forecasting curve for the recently flaring active regions, and another forecasting curve for the active regions that have not recently flared. If the two forecasting curves are essentially the same, this implies that the secondary parameter does not provide additional information that improves the performance over that of the forecasting technique that is based on only the primary parameter. When the two forecasting curves differ by a statistically significant amount, this indicates that the two-parameters technique performs better than the corresponding single-parameter technique. Depending on the relative sample size and the relative number of events, the number of bins used might be different for the two forecasting curves, as is the case in the present example.

The plot on the right in Figure 2 shows the two forecasting curves for Next MAG4, the forecasting technique that uses both free-energy proxy ${}^{L}WL_{SG}$ and previous X- and M-class flare history (orange is for previous-flaring ARs, and black is for no-previous-flaring ARs). The plot shows the two groups definitely have different dependence on the free-energy proxy. For active regions with the largest free-energy proxy, both groups have approximately the same expected

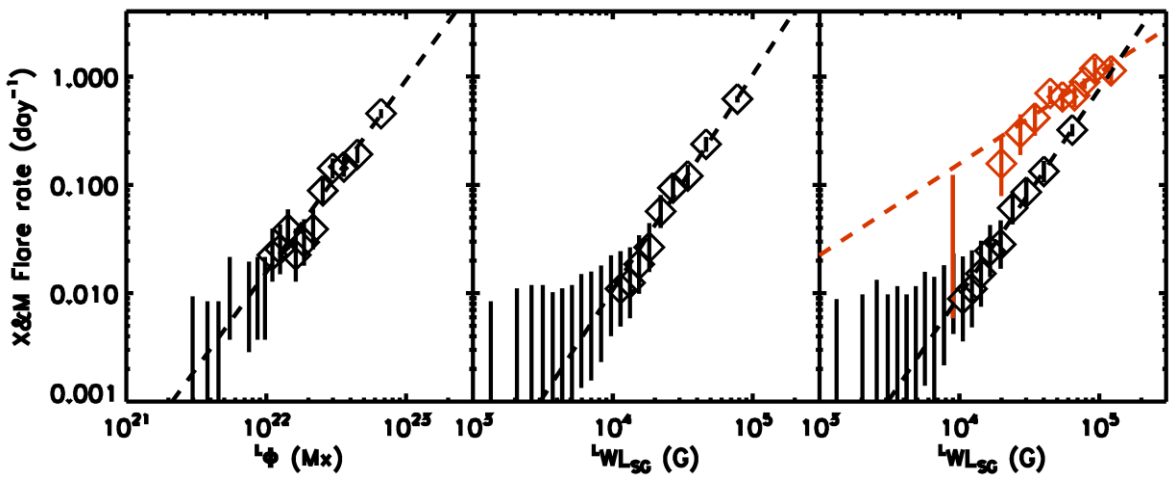

Figure 2: Empirically derived forecasting curves for three forecasting techniques**. Left:** Forecasting curve for forecasting based on total magnetic flux. **Middle:** Forecasting curve for forecasting based on free-energy proxy only. **Right:** Forecasting curves for forecasting based on free-energy proxy combined with previous flare activity (black: for ARs that have had no recent X- or M-class flares, orange: for ARs that have had recent X- or M-class flares). The 40,000 magnetograms are sorted by either total magnetic flux (left) or free-energy proxy (center and right), and then binned. A diamond is centered on the average measure and average event rate in each bin having total magnetic flux ${}^{L}\Phi > 10^{22}$ Mx or free-energy proxy ${}^{L}WL_{SG} > 10^{4}$ G. The vertical line in each bin is the 1-σ uncertainty in the event rate based on Poisson statistics. Each dashed line is a least-squares linear fit through the diamonds.

X- and M-class event rate (here nearly all of the ARs are previous-flaring ARs). We note that for active regions with moderately-large ${}^{L}WL_{SG}$ (~3 x 10^{4} G) the forecasted flare rate for previous-flaring ARs is about 0.3 flares/day greater than for no-previous-flaring ARs. For active regions with average ${}^{L}WL_{SG}$ (~10^{4} G), the forecasted flare rate for previous-flaring ARs is increased by a factor of 14 over that for no-previous-flaring ARs. For active regions with ${}^{L}WL_{SG}$ below average (<10^{4} G), the forecasted flare rates for previous-flaring active regions are poorly determined due to small sample size (these active regions are so small that they rarely produce a major flare, X- or M- class).

Forecasted event rate can be converted into forecast event probability assuming Poisson statistics (Wheatland 2004). The probability of an event is $(1\text{-}e^{-R\Delta T})$, where ΔT is the temporal window for the forecast and R is the forecasted event rate. It should be noted that full-disk event rates, when there are multiple active regions on the disk and assuming no correlations, are additive, while probabilities are not. This is why MAG4 forecasts the full-disk event rate and converts that into a full-disk event probability.

4. Comparison of Average Values of the Metrics for the Four Forecasting Techniques

Figure 2 implies that the forecasting technique based on both free-energy proxy and previous flare activity (Next MAG4) should perform significantly better than the technique based on the free-energy proxy alone (Present MAG4). In this section we measure how much Next MAG4 outperforms Present MAG4, and how much Present MAG4 outperforms the two techniques based on either total magnetic flux or McIntosh AR class. Our approach is to randomly divide the ARs in our database into two sets, a control set and an experimental set,

Table 1. Two-By-Two Contingency Table Format and Definitions of our Forecasting-Performance Metrics

	Event Observed	
Event Forecast	Yes	No
Yes	A	B
No	C	D

Percent Correct, PC = (A+D)/N
Number of Forecasts, N = A+B+C+D
Probability of Detection, POD = A/(A+C)
False Alarm Rate, FAR=B/(A+B)
Heidke Skill Score, HSS = (A+D-E)/(N-E)
E = ((A+B)(A+C)+(B+D)(C+D))/N
True Skill Score, TSS=(AD-BC)/[(A+C)(B+D)]

each having half of the ARs in the database. For each division, we obtain the forecasting curve from the control half and apply this forecasting curve to the experimental half. To measure whether one forecasting technique is significantly better than another, we use the repeated random sub-sampling cross-validation method (Shao 1993). To measure the performance of the McIntosh AR Class technique, we have to modify the above approach that we use for the other three forecasting techniques, because some McIntosh AR classes are so rare that the control subsample often has a statistically insignificant number of these ARs while the experimental subsample has some ARs of this class. So for each random division of the ARs in the database into control and experimental subsamples, we don't use the control subsample in obtaining the X- and M- class flare rates forecast for the ARs in the experimental subsample by the McIntosh AR Class technique. Instead, for each AR in the experimental subsample, for its flare rate forecast by McIntosh AR Class, we take the average flare rate of ARs of the McIntosh class of that AR. This flare rate is given for that class in the NOAA Table. To ensure that the control and experimental subsamples are independent of each other, we randomly assign active regions, not individual AR magnetograms, to either the control or experimental subsample. In other words, when we divide the sample a random number is assigned to each active region instead of to each active-region magnetogram, so that all magnetograms from the same active region are assigned either to the control subsample or to the experimental subsample.

From each Monte-Carlo run we obtain a truth table of the format of Table 1 for each forecasting technique. We do this by obtaining the forecasting curves for the Total Magnetic Flux technique, the Present MAG4 technique, and Next MAG4 technique from the control set of ARs. We apply these forecasting curves to the experimental set to obtain for each AR magnetogram in the experimental set the forecasted flare rate for each of these three techniques. We use the NOAA flare rate table to obtain for each AR in the experimental set the forecasted flare rate for the McIntosh AR Class technique. For each AR magnetogram giving a forecasted event rate of 0.5 major flares or greater during the forecast window (24 hours for this study), we forecast the active region will produce a major flare during the forecast window (a Yes in the truth table), and for all cases where the forecasted event rate is less than 0.5, we forecast the active region will not produce a major flare (a No in the truth table). Also, counted in the truth table is whether the active region did or did not produce a major flare in the 24 hours after the time of the magnetogram. From the contingency table (the truth table), we evaluate five different technique-performance metrics (see Balch 2008): these performance metrics are Percent

Correct (PC), Probability of Detection (POD), False Alarm Rate (FAR), Heidke Skill Score (HSS), and True-Skill Score (TSS), defined in Table 1. The PC is always high since most active regions (~95%) do not produce a major flare in the next 24 hours. The POD is the fraction of the major flares that were forecasted; lowering the threshold below 0.5 would improve the POD score. The FAR is the fraction of AR magnetograms from which a major flare is forecasted for but no flare occurs. Raising the threshold for Yes forecasts would lower the average value of the FAR metric, an improvement in the performance measured by the FAR metric. The HSS is similar to the PC, but corrects for climatology (corrects for the outcome being No in 95% of the trials). The TSS is POD minus the ratio B/(B + D) (Woodcock 1976). The ratio B/(B + D) is the number of false alarms (B) divided by the total number of No outcomes (B + D). Since each of the five metrics has its biases, evaluating all five determines with more confidence whether one forecasting technique performs significantly better than the others.

For each of the 2000 Monte-Carlo runs, we determine $M_{ij}^{(k)}$, where M stands for metric, i enumerates the run, j enumerates the metric, and k enumerates the forecasting technique. In Table 2, we list the average, $M_j^{(k)}$ (= $\sum_i M_{ij}^{(k)}/N$, where N is the total number of runs), and the standard deviation ($\sigma_j^{(k)}$) of the distribution of $M_{ij}^{(k)}$ from all the runs. For each entry in Table 2, the value in the parenthesis is the statistical significance of the difference between the average metric of Present MAG4 and the average metric of the technique for that entry, and is defined as:

$$\delta_j^{(k1,k2)} = (M_j^{(k1)} - M_j^{(k2)})/[(\sigma_j^{(k1)})^2 + (\sigma_j^{(k2)})^2]^{0.5}. \quad (6)$$

In Equation (6), the index k2 refers to the Present MAG4 technique. The distributions ($M_{ij}^{(k)}$), the average ($M_j^{(k)}$), and standard deviation ($\sigma_j^{(k)}$) for each forecasting technique and each metric are shown in Figure 3. In Equation (6), it is appropriate to use the standard deviations of the two distributions instead of the standard deviations of the two means, because when the difference between two means is much less than the combined standard deviation of the two distributions (the denominator in Equation (6)), in almost half random-division runs the difference in the metric between the two techniques is in the opposite direction to that of the difference in the two means.

The results from Equation (6) are given in parentheses in Table 2. Table 2 is read as follows, using PC for McIntosh AR Class for example: 93.7 ± 0.7 (-2.2). Here 93.7 is the average value of PC, 0.7 is the standard deviation of the PC distribution from 2000 runs for McIntosh AR Class, and 2.2 is the statistical significance (given by Equation 6) of the difference of this average from the Present MAG4 average PC, with the minus sign indicating that McIntosh AR Class performs worse than Present MAG4.

Figure 3 and Table 2 confirm quantitatively that Next MAG4 performs better than Present MAG4. They also show that Present MAG4 performs significantly better than either Total Magnetic Flux or McIntosh AR Class. The latter two techniques have comparable performance with McIntosh AR Class being worse in 3 metrics and better in two metrics than Total Magnetic Flux. When we compare the significances of the differences in the means of the metrics, McIntosh AR Class compared to Present MAG4 performs worse by 2 standard deviations in three metrics (PC, FAR, and TSS), while Total Magnetic Flux compared to Present MAG4 performs worse by 1.8 standard deviations in two metrics (POD and HSS). Next MAG4 performs better in all five metrics relative to Present MAG4, but never by more than 1 standard deviation.

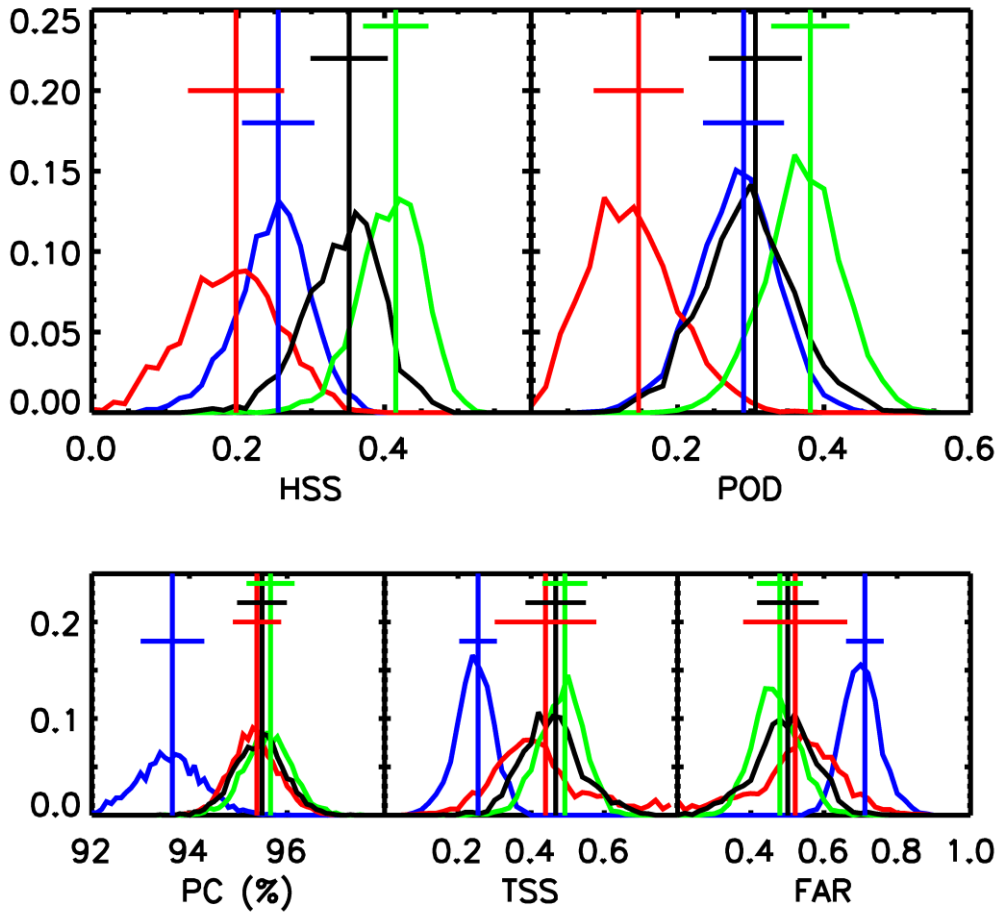

Figure 3 Comparison of the performance of the four forecasting techniques in terms of their distributions and average values of each of five performance metrics. Each panel is for a different performance metric. The distribution of the measured metric for 2000 runs is shown for each forecasting technique: McIntosh AR Class (blue), Total Magnetic Flux (red), Present MAG4 (black) and Next MAG4 (green). The vertical line shows the average of the distribution, while the horizontal line shows the standard deviation of the distribution. For each metric, the McIntosh AR Class technique and the Total Magnetic Flux technique perform worst (note that a lower value is better for FAR). In all cases Next MAG4 performs best, with Present MAG4 in second place in average value of the metric. For some metrics the difference in performance between some of the techniques is not statistically significant.

Other thresholds can be used to convert forecasted event rate (or event probability) into yes/no forecasts as in Table 1. We have confirmed that a 50% probability threshold (event rate of 0.69), and a 0.4 event rate ranks the forecasting techniques in the same order, indicating a robustness of the result from using an event rate of 0.5 for the yes/no discriminator.

5. Comparison of Same-Metric Differences for the Four Forecasting Techniques

While Table 2 ranks the four techniques roughly in the order we expect, the significance of the difference is low. Except for McIntosh AR Class versus Present MAG4 the results do not appear to be statistically strongly significant. Even though some individual metrics results are not statistically significant, every metric difference is in the same direction, so there appears to be a systematic trend. This indicates that the spread of the distribution is hiding a significant systematic effect. When we compare metrics for Present MAG4 to metrics for McIntosh AR Class, we find that in all 2000 runs Present MAG4 had a higher PC, a higher TSS and a better (lower) FAR; in only 29 (~1.5% of the runs) runs McIntosh AR Class had a higher HSS, and in only 752 (15% of the cases) a higher POD. This is obviously better than a ~2 σ difference in performance. We have found that the metrics from these two techniques, and from the other pairs of techniques, have some correlation. That is, from runs where one technique gets a lower than average value in a metric, each the other three forecasting techniques also tends to get a lower than average value in the same metric. This means that for each metric there is some correlation in the metric's values for each pair of forecasting techniques. To remove this correlation, we need to compare the average difference as well as the standard deviation of the average difference in the metrics between two forecasting techniques. The average difference of metric j is

$$\Delta M_j^{(k1,k2)} = \sum_i (M_{i,j}^{(k1)} - M_{i,j}^{(k2)})/N, \quad (7)$$

where k1 and k2 refer to the two different forecasting techniques being compared. In Table 3, k2 refers to Present MAG4. For each of the other three techniques compared to Present MAG4, Table 3 gives for each metric the average difference of the metric, the standard deviation of the distribution of the differences, and the ratio of the average difference to the standard deviation of the distribution of the differences. The ratio of the mean difference to the standard deviation of the distribution of the differences is the number of sigma of the mean difference from 0, the statistical significance of the mean difference.

Figure 4 and Table 3 show from the same-run metric differences that Present MAG4 performs better than McIntosh AR Class in all five metrics, and significantly so in four of the five metrics (PC by 4 σ, FAR by 3.1 σ, HSS by 2.2 σ, and TSS by 3.1 σ). Present MAG4 performs better than Total Magnetic Flux in all five metrics, and significantly in two of the five metrics (POD by 2.4 σ, and HSS by 2.2 σ). Next MAG4 performs better than Present MAG4 in all five metrics, and significantly so in two of the five metrics (POD by 2.3 σ, and HSS by 2.3 σ). In addition to the results in Table 3, we also found that the performance of Next MAG4 relative

Table 2 Comparison of the Performance of the other Three Forecasting Techniques to that of Present MAG4 in Terms of Average Values of the Performance Metrics

Technique	PC	POD	FAR	HSS	TSS
McIntosh AR Class	93.7 ± 0.7 (-2.2)	0.29 ± 0.06 (-0.19)	0.71 ± 0.05 (-2.1)	0.25 ± 0.05 (-1.3)	0.26 ± 0.05 (-2.2)
Total Magnetic Flux	95.4 ± 0.5 (-0.15)	0.15 ± 0.06 (-1.8)	0.52 ± 0.14 (-0.12)	0.20 ± 0.07 (-1.8)	0.44 ± 0.14 (-0.17)
Present MAG4	95.5 ± 0.5 (NA)	0.31 ± 0.06 (NA)	0.50 ± 0.08 (NA)	0.35 ± 0.05 (NA)	0.47 ± 0.08 (NA)
Next MAG4	95.7 ± 0.5 (0.24)	0.38 ± 0.05 (0.91)	0.48 ± 0.06 (0.21)	0.42 ± 0.04 (0.92)	0.49 ± 0.06 (0.24)

Notes: the lower FAR the better the forecast.
The value in the parenthesis is the statistical significance (number of sigma) of the difference in the average metric between the listed forecast technique and Present MAG4, given by Equation (6).

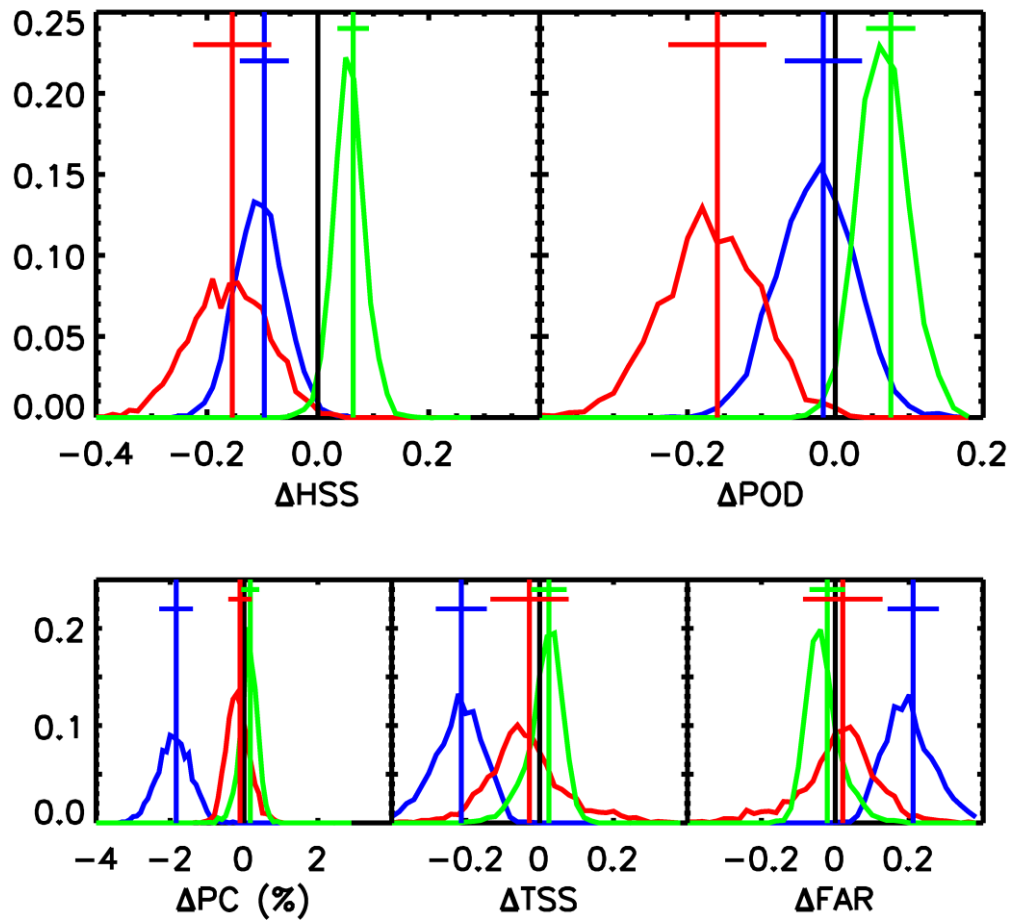

Figure 4 Comparison of the distributions and average values of the same-run differences between each performance metric for MAG4 and the same performance metric for each of the other three forecasting techniques. Each metric-difference distribution from 2000 runs is shown in the five panels, each panel for a separate metric. In each panel, McIntosh AR Class (blue), Total Magnetic Flux (red), and Next MAG4 (green) are compared to Present MAG4. The blue, red, and green vertical lines each show the average of the difference of that metric, while the horizontal line shows the standard deviation of the distribution. For any given metric, a technique that performed equally well as MAG4 would have its metric-difference mean at 0. Except for ΔFAR, a technique that performed significantly worse (or better) than Present MAG4 has a metric-difference mean that is more than one standard deviation to the left (or right) of 0. A mean farther to the left is better for ΔFAR. Next MAG4 makes more accurate forecasts than Present MAG4, which in turn is more accurate than Total Magnetic Flux or McIntosh AR Class.

to McIntosh AR Class is significantly better in all five metrics (PC by 4.3 σ, POD by 1.9 σ, FAR by 4.2 σ, HSS by 3.9 σ, and TSS by 4.3 σ). These results establish quantitatively that Present MAG4 is superior to either McIntosh AR Class or Total Magnetic Flux, and confirm that Next MAG4 is superior to Present MAG4.

6. Conclusions and Discussion

We have quantitatively demonstrated, using 2x2 contingency tables and five performance metrics, that for active regions within 30 heliocentric degrees of disk center forecasting of an active region's rate of X- and M-class flares is significantly more accurate when based on our free-energy proxy (Present MAG4 technique) than when based on McIntosh AR classes (the

Table 3 Comparison of the Performance of the Other Three Forecasting Techniques to that of Present MAG4 in Terms of Average Same-Run Same-Metric Differences

Other Forecasting Technique	ΔPC(%)	ΔPOD	ΔFAR	ΔHSS	ΔTSS
McIntosh AR Class	-1.8 ±0.5 (-4.0)	-0.02 ± 0.05 (-0.31)	-0.21 ± 0.07 (-3.1)	-0.10 ± 0.04 (-2.2)	-0.21 ± 0.07 (-3.1)
Total Magnetic Flux	-0.1 ± 0.3 (-0.33)	-0.16 ± 0.07 (-2.4)	-0.02 ± 0.11 (-0.19)	-0.15 ± 0.07 (-2.2)	-0.03 ± 0.11 (-0.26)
Next MAG4	0.2 ± 0.2 (0.72)	0.08 ± 0.03 (2.2)	0.02 ± 0.05 (0.46)	0.06 ± 0.03 (2.3)	0.03 ± 0.05 (0.53)

Notes: Value in parentheses is the mean difference divided by the variance of the distribution of the difference using sign convention of positive for improvement over present MAG4. We have reversed the sign of ΔFAR in order for it to be the same as the other 4 metrics, in that a positive value indicates the forecast is better than free-energy proxy only, and negative indicates the forecast is worse. Note we have not reversed the sign of ΔFAR in Figure 4. ΔPC etc stand for $\Delta M_j^{(k1,k2)}$, comparing the performance of the forecasting technique in the left column to that of present MAG4.

starting point used by NOAA, but not their final product) or than when based on total magnetic flux. We have further demonstrated that forecasting based on previous flare history combined with our free-energy proxy (Next MAG4) is superior to forecasting based on free-energy proxy only (Present MAG4). We plan to upgrade the MAG4 tool to Next MAG4 in the coming year. Currently MAG4 makes forecasts beyond 30 heliocentric degrees and warns of reduced accuracy for these forecasts for ARs beyond 30^o. This limit will be lessened when MAG4 begins using HMI vector magnetograms. We have to deproject the HMI vector magnetograms and measure WL_{SG} so that we can use the huge MDI database, which is the only adequate database so far.

We have further found that the performance of the Next MAG4 technique is insensitive to both the temporal window used to define whether an active region has recently been flare productive and the temporal window used to obtain forecasted rates from the database (Appendix A). While we will use 24 hours forward and 24 hours backward for Next MAG4, either or both of these windows could be 48 hours for essentially the same performance.

We also investigated whether forecasting from our free-energy proxy combined with any McIntosh AR broad category or any Mount Wilson AR class could improve the performance over that of forecasting from free-energy proxy only. We found no significant improvement (Appendix B).

We have demonstrated a method of measuring whether one forecasting technique is significantly better than an alternative technique. This method needs a large-enough sample of ARs to work well since smaller samples have larger variations in the control-sample forecasting curves and thus larger variations in a metric difference between two techniques, and hence have less resolution for ranking the performance of the two forecasting techniques. We also note that our method of measuring the difference in performance between comparable forecasting techniques works well only when the performance of the techniques is compared for forecasting of events for which at least some active regions have expected event rates above the threshold for an AR to have a Yes forecast in the 2x2 contingency table (greater than 0.5 events per length of the forward window used to obtain the AR's expected rate from the control sample of ARs). For rare events, such as SPEs of X10+ flares, the expected event rate is always less than 0.5 (for evaluation windows of less than four days), the forecast for an AR is always No in the contingency table, and the performance metrics used in this paper are insensitive to the difference in performance between alternative forecasting techniques.

Table 4: Next MAG4 Performance Improvement in Terms of Average Same-Run Metric Differences for Different Temporal Windows Compared to "24/24"

"24/24" to	ΔPC(%)	ΔPOD	ΔFAR	ΔHSS	ΔTSS
"24/48"	-0.1 ± 0.2 (-0.31)	0.01 ± 0.02 (0.29)	-0.01 ± 0.03 (-0.31)	0.00 ± 0.02 (0.02)	-0.01 ± 0.03 (-0.30)
"24/72"	0.0 ± 0.2 (0.02)	0.00 ± 0.03 (-0.09)	-0.00 ± 0.03 (0.05)	0.00 ± 0.02 (-0.09)	0.00 ± 0.03 (0.05)
"24/96"	-0.1 ± 0.2 (-0.27)	-0.02 ± 0.04 (-0.48)	-0.01 ± 0.04 (-0.24)	-0.02 ± 0.03 (-0.56)	-0.01 ± 0.04 (-0.26)
"48/24"	0.1 ± 0.1 (0.70)	-0.02 ± 0.02 (-0.94)	0.02 ± 0.02 (0.86)	-0.01 ± 0.02 (-0.53)	0.02 ± 0.02 (0.82)
"48/48"	0.1 ± 0.2 (0.26)	-0.02 ± 0.03 (-0.69)	0.01 ± 0.03 (0.30)	-0.01 ± 0.03 (-0.45)	0.01 ± 0.03 (0.27)
"48/72"	0.1 ± 0.2 (0.39)	-0.03 ± 0.04 (-0.80)	0.02 ± 0.04 (0.44)	-0.01 ± 0.03 (-0.46)	0.02 ± 0.04 (0.41)
"48/96"	0.0 ± 0.3 (0.02)	-0.05 ± 0.04 (-1.24)	-0.00 ± 0.04 (0.07)	-0.03 ± 0.03 (-0.98)	0.00 ± 0.04 (0.02)
"72/24"	0.1 ± 0.2 (0.35)	-0.04 ± 0.04 (-1.1)	0.02 ± 0.03 (0.47)	-0.02 ± 0.03 (-0.87)	0.01 ± 0.03 (0.43)
"72/48"	0.1 ± 0.2 (0.35)	-0.04 ± 0.04 (-1.0)	0.02 ± 0.04 (0.46)	-0.03 ± 0.03 (-0.75)	0.02 ± 0.04 (0.42)
"72/72"	0.1 ± 0.3 (0.47)	-0.05 ± 0.04 (-1.1)	0.03 ± 0.05 (0.56)	-0.02 ± 0.03 (-0.75)	0.02 ± 0.05 (0.52)
"72/96"	0.0 ± 0.3 (0.14)	-0.06 ± 0.04 (-1.5)	0.01 ± 0.05 (0.23)	-0.04 ± 0.03 (-1.2)	0.01 ± 0.05 (0.18)
"96/24"	0.0 ± 0.3 (0.13)	-0.05 ± 0.04 (-1.2)	0.01 ± 0.05 (0.28)	-0.03 ± 0.03 (-1.1)	0.01 ± 0.05 (0.24)
"96/48"	0.1 ± 0.3 (0.34)	-0.06 ± 0.05 (-1.1)	0.02 ± 0.05 (0.47)	-0.03 ± 0.04 (-0.91)	0.02 ± 0.05 (0.43)
"96/72"	0.1 ± 0.3 (0.48)	-0.06 ± 0.04 (-1.3)	0.03 ± 0.05 (0.59)	-0.03 ± 0.04 (-0.96)	0.03 ± 0.05 (0.55)
"96/96"	0.1 ± 0.3 (0.23)	-0.07 ± 0.04 (-1.7)	0.02 ± 0.05 (0.35)	-0.05 ± 0.03 (-1.4)	0.01 ± 0.05 (0.29)

Notes: Value in parentheses is the statistical significance (in sigma) of the mean difference, with sign convention of positive (negative) for better (worse) performance compared to the (24-hour-backward /24-hour-forward) window pair.

ΔFAR sign convention: positive is for better performance than "24/24", negative is for worse

Appendix A: Sensitivity of Performance of Next MAG4 to Length of Temporal Windows

We have shown that forecasting based on free-energy proxy combined with previous flare history (Next MAG4) is superior to the other three methods considered so far in this paper. Next MAG4 forecasting is based on whether the active region has produced an X- or M-class flare in the previous 24 hours, and on obtaining the forecasting curves from the observed AR flare production in the coming 24 hours in the database. These periods were chosen partially due to the length of a day on Earth, and thus, from the Sun's point of view, are arbitrary. In this Appendix, we address two questions: 1) Are there some other backward and forward windows that improve the performance of Next MAG4, and 2) how significant is the improvement for a particular new window. We measured the difference in performance between backward windows of 24, 48, 72, and 96 hours, and forward windows of 24, 48, 72, and 96 hours. Each (backward/forward) pair of windows constitutes a different forecasting technique. In Table 4, in the same way as in Table 3, each forecasting technique's performance is compared, to that of the "24/24" window-pair "standard" technique. In all cases, to evaluate the different metrics, we use in the contingency tables the outcomes in a forward window of 24 hours. That this evaluation window is the same for evaluations of each technique is required to avoid inducing biases. Shorter evaluation windows would result in fewer Yes forecasts, and fewer Yes forecasts would result in better scores for PC and FAR, with the opposite effect for POD. If the duration of the evaluation window were different for two compared techniques, these biases would bias the measured difference in performance.

As can be seen in Table 4, there is no case in which all five metric-difference means are positive. Typically ΔPC, ΔFAR, and ΔTSS are positive, and ΔPOD and ΔHSS have negative means (or the reverse for "24/48" hours). For "24/96" hours compared to "24/24" hours the performance is always worse. For "48/24", ΔPC, ΔFAR and ΔTSS have greater than ~0.7 σ significance, but ΔPOD and ΔHSS means are worse with significances of 0.5 σ or greater.

Longer backward windows (72 and 96 hours) have ΔPOD means worse by more than 1 σ, and have ΔHSS means worse by ~ 1 σ. These results indicate that there is no window pair that is significantly better than "24/24" hours. Generally, the metric-difference means between the different-window-pair techniques are smaller than the metric-difference means between Present MAG4 and Next MAG4 (see Tables 3 and 4).

Appendix B: Performance of Techniques Based on Mount Wilson or McIntosh AR Class Combined with Free-Energy Proxy

Using the method described in this paper we can investigate whether any of the multiple factors that decide which McIntosh AR class to assign an active region to can be used to improve forecasting based on free-energy proxy. This is done in the same manner as for forecasting from free-energy proxy and previous flare activity; we divide the sample into those active regions with a given attribute and those without. For example, we can divide the entire sample based on whether the AR is classified as compact or not by McIntosh. An active region is classified as compact if the third letter in the AR's McIntosh AR class is C. All active regions with the third letter being C are assigned to one group (just like recent flare activity), and all active regions with the third letter not being C are assigned to a second group (just like no recent flare activity). For each group we derive separate forecasting curves (as in the right panel of Figure 2). The broad categories of McIntosh AR classes that we tested are listed in the first column of Table 5. Each broad category is specified by a letter that, together with the letter's position in the 3-letter sequence of the McIntosh AR class designations, stands for an AR attribute. The ARs of any one broad category all have the same letter in the same one of the three positions and have any possible letters in the other two positions (denoted by asterisks in Table B1). The second column in Table 5 briefly states what AR attribute the letter and its position represent (See McIntosh 1990, Bornmann, Kalmbach, and Kulhanek 1994, Bornmann and Shaw 1994 for a more detailed description of the classes).

We also tried the Mount Wilson active-region classification system (Zirin 1988, Hale et al. 1919). In the Mount Wilson system an active region is either an α or a β active region. Active regions that are α active regions have sunspots of only one polarity; active regions that are β active regions have sunspots of both polarities. Active regions with both polarities intermixed are classified as γ active regions. If the active region has two opposite polarity sunspots that share the same penumbra, its classified as a δ active region. Hence each active region is one of the following five classes: α, β, βγ, βδ, or βγδ. We checked whether there were any significant improvements in performance over Present MAG4 when combining any of these AR classes with free-energy proxy, and found none.

The results indicate that none of the broad categories of McIntosh AR classes and none of Mount Wilson AR classes give any significant improvement in forecasting performance; none of the metrics show an improvement in performance of more than 1 standard deviation in the metric-difference means. Also, there is no AR McIntosh broad category or Mount Wilson class that gives at least some improvement across all five metrics. This is in contrast to forecasting based on free-energy proxy in combination with flare history (Next MAG4) compared to Present MAG4, where the metric-difference means show improvement for all five metrics, and for two of these, by more than twice the standard deviation in the metric-difference mean. This indicates that combining free-energy proxy with any McIntosh AR broad category or Mount Wilson AR class does not give a significant improvement in the performance over that of Present MAG4. These results also demonstrate that examining the improvement in performance of a technique

Table 5: Performance Improvement Relative to Present MAG4 from Forecasting from a McIntosh AR Broad Category Combined with AR Free-Energy Proxy

	Broad Category	ΔPC(%)	ΔPOD	ΔFAR	ΔHSS	ΔTSS
A**	Unipolar without penumbra	0.0 ± 0.0 (0.05)	0.00 ± 0.00 (-0.04)	0.00 ± 0.00 (0.00)	0.00 ± 0.00 (-0.02)	0.00 ± 0.00 (0.00)
B**	Bipolar/no penumbra	0.0 ± 0.0 (0.07)	0.00 ± 0.00 (-0.38)	0.00 ± 0.01 (0.03)	0.00 ± 0.00 (-0.27)	0.00 ± 0.01 (0.02)
C**	Bipolar, 1 penumbra	-0.1 ± 0.3 (-0.26)	0.00 ± 0.00 (-0.78)	-0.01 ± 0.03 (-0.38)	-0.01 ± 0.01 (-0.59)	-0.01 ± 0.03 (-0.39)
D**	Bipolar, Penumbra both ends Longitudinal extent <10°	-0.2 ± 0.2 (-0.83)	0.00 ± 0.03 (0.16)	-0.03 ± 0.03 (-0.78)	-0.01 ± 0.02 (-0.26)	-0.03 ± 0.03 (-0.77)
E**	Bipolar, Penumbra both ends Longitudinal extent 10° to 15°	-0.1 ± 0.1 (-0.53)	0.00 ± 0.02 (-0.09)	-0.02 ± 0.03 (-0.56)	-0.01 ± 0.02 (-0.29)	-0.02 ± 0.03 (-0.55)
F**	Bipolar, Penumbra both ends Longitudinal extent > 15°	-0.1 ± 0.1 (-0.63)	0.00 ± 0.03 (-0.05)	-0.01 ± 0.03 (-0.58)	-0.01 ± 0.03 (-0.25)	-0.01 ± 0.03 (-0.58)
H**	Unipolar with penumbra	0.0 ± 0.0 (0.52)	0.00 ± 0.00 (-0.58)	0.00 ± 0.00 (0.49)	0.00 ± 0.00 (-0.32)	0.00 ± 0.00 (0.48)
A	Largest spot has mature asymmetric penumbra <2.5° north-south diameter	-0.3 ± 0.2 (-1.3)	0.01 ± 0.02 (0.39)	-0.04 ± 0.04 (-1.1)	-0.01 ± 0.02 (-0.40)	-0.04 ± 0.04 (-1.1)
H	Largest spot has mature symmetric penumbra >2.5° north-south diameter	-0.1 ± 0.3 (-0.30)	0.00 ± 0.01 (-0.04)	-0.01 ± 0.04 (-0.32)	0.00 ± 0.01 (-0.34)	-0.01 ± 0.04 (-0.33)
K	Largest spot has mature asymmetric penumbra >2.5° north-south diameter	-0.1 ± 0.2 (-0.75)	0.00 ± 0.02 (0.15)	-0.02 ± 0.03 (-0.66)	-0.01 ± 0.02 (-0.29)	-0.02 ± 0.03 (-0.65)
R	Rudimentary penumbra surrounds largest spot	-0.1 ± 0.4 (-0.30)	0.00 ± 0.00 (-0.22)	-0.02 ± 0.05 (-0.32)	-0.01 ± 0.02 (-0.34)	-0.02 ± 0.05 (-0.32)
S	Largest spot has mature symmetric penumbra <2.5° north-south diameter	0.0 ± 0.5 (-0.04)	0.00 ± 0.01 (-0.31)	-0.00 ± 0.02 (-0.03)	0.00 ± 0.01 (-0.18)	0.00 ± 0.02 (-0.03)
X	AX* or BX*	0.0 ± 0.0 (0.26)	0.00 ± 0.00 (-0.40)	0.00 ± 0.00 (0.21)	0.00 ± 0.00 (-0.23)	0.00 ± 0.00 (0.20)
**C	Compact	-0.1 ± 0.1 (-0.72)	0.01 ± 0.02 (0.50)	-0.02 ± 0.03 (-0.63)	0.00 ± 0.02 (0.12)	-0.02 ± 0.03 (-0.61)
**I	Intermediate	-0.1 ± 0.1 (-0.82)	0.00 ± 0.01 (0.22)	-0.01 ± 0.02 (-0.65)	0.00 ± 0.01 (-0.13)	-0.01 ± 0.02 (-0.63)
**O	Open	-0.1 ± 0.2 (-0.41)	0.01 ± 0.02 (0.28)	-0.01 ± 0.03 (-0.40)	0.00 ± 0.02 (-0.01)	-0.01 ± 0.03 (-0.40)
**X	A*X or H*X	0.0 ± 0.0 (0.55)	0.00 ± 0.00 (-0.56)	0.00 ± 0.00 (0.50)	0.00 ± 0.00 (-0.29)	0.00 ± 0.00 (0.49)
A	Unipolar	0.0 ± 0.0 (0.54)	0.00 ± 0.00 (-0.56)	0.00 ± 0.00(0.49)	0.00 ± 0.00 (-0.29)	0.00 ± 0.00 (0.48)
BD	Delta sunspot	-0.1 ± 0.1 (-0.61)	0.00 ± 0.02 (0.29)	-0.01 ± 0.02 (-0.57)	0.00 ± 0.01 (-0.08)	-0.01 ± 0.02 (-0.56)
BG	Mixed polarity	0.0 ± 0.1 (-0.34)	0.00 ± 0.01 (-0.32)	-0.00 ± 0.02 (-0.27)	0.00 ± 0.01 (-0.44)	0.00 ± 0.02 (-0.28)
BGD	Mixed polarity with delta sunspot	-0.1 ± 0.1 (-0.46)	0.00 ± 0.02 (0.14)	-0.01 ± 0.02 (-0.41)	0.00 ± 0.01 (-0.14)	-0.01 ± 0.02 (-0.40)

for each of several metrics rather than for only a single metric helps decide whether one forecasting technique is significantly better than another.

We conclude that most of the factors that allow the McIntosh AR classes to be used to make flare forecasts (i.e., large, compact, asymmetric penumbra) are correlated with a large free-energy proxy. Thus, while predictive by themselves, they do not seem to include any relevant additional information, unlike previous flare activity.

Acknowledgements. Support for MAG4 development comes from NASA's Game Changing Development Program, Johnson Space Center's Space Radiation Analysis Group (SRAG), and AFOSR's Multi-University Research Initiative. In particular, the authors want to gratefully acknowledge the continued support and guidance of Dr. Dan Fry (NASA-JSC) and David Moore (NASA-LaRC). The flare information used in this study was provided by NOAA /SWPC. We acknowledge the use of the SOHO LASCO CME Catalog. This CME catalog is generated and maintained at the CDAW Data Center by NASA and The Catholic University of America in cooperation with the Naval Research Laboratory. SOHO is a project of international cooperation between ESA and NASA. We also used data supplied courtesy of SolarMonitor.org. The authors also want to gratefully acknowledge the useful and insightful statistics discussions with Dr. Robert Wilson (NASA-MSFC). We wish to thank the reviewers for useful comments that improved the paper.

References

Alissandrakis, C. E. (1981), On the Computation of Constant Alpha Force-Free Magnetic Field, *Astronomy and Astrophysics, 100,* 197.

Abramenko, V. I. (2005), Relationship between Magnetic Power Spectrum and Flare Productivity in Solar Active Regions, *Astrophys. J.,, 629,* 1141

Balch, C. C. (2008), Updated Verification of the Space Weather Prediction Center's Solar Energetic Particle Prediction Model, *Space Weather, 6,* S01001

Barnes, G., & Leka, K. D. (2008), Evaluating the Performance of Solar Flare Forecasting Methods, *Astrophys. JL, 688,* L107

Bornmann, P. L., Kalmbach, D., Kulhanek, D., (1994), McIntosh active-region class similarities and suggestions for merger, *Sol. Phys. 150,* 147

Bornmann, P. L., Shaw D., (1994), Flare Rates and the McIntosh Active-Region Classifications, *Sol Phys. 150,* 127

Cui, Y., Li, R., Zhang, L., He, Y., & Wang, H. (2006), Correlation between Solar Flare Productivity and Photospheric Magnetic Field Properties, *Sol. Phys., 237,* 45

Falconer, D. A., (2001), A Prospective Method for Predicting Coronal Mass Ejections from Vector Magneotgrams, J. Geophys. Res. 106 25,185.

Falconer, D. A., Barghouty, A. F., Khazanov, I., & Moore, R. (2011), A Tool for Empirical Forecasting of Major Flares, Coronal Mass Ejections, and Solar Particle Events from a Proxy of Active-Region Free Magnetic Energy, *Space Weather, 9,* S04003

Falconer, D. A., Moore, R. L., Barghouty, A. F., & Khazanov, I. (2012), Prior Flaring as a Complement to Free Magnetic Energy for Forecasting Solar Eruptions, *Astrophys. J., 757,* 32

Falconer, D. A., Moore, R. L., & Gary, G. A. (2002), Correlation of the Coronal Mass Ejection Productivity of Solar Active Regions with Measures of their Global Nonpotentiality from Vector Magnetograms: Baseline Results, *Astrophys. J, 569,* 1016

Falconer, D. A., Moore, R. L., & Gary, G. A. (2003), A Measure from Line-of-Sight Magnetograms for Predicting Coronal Mass Ejections, *J. Geophys. Res., 108* (A10), 1380

Falconer, D. A., Moore, R. L., & Gary, G. A. (2006), Magnetic Causes of Solar Coronal Mass Ejections: Dominance of the Free Magnetic Energy of the Magnetic Twist Alone, *Astrophys. J., 644,* 1256

Falconer, D. A., Moore, R. L., & Gary, G. A. (2008), Magnetogram Measures of Total Nonpotentiality for Prediction of Solar Coronal Mass Ejections from Active Regions of Any Degree of Magnetic Complexity, Astrophys. *J., 689,* 1433

Falconer, D. A., Moore, R. L., Gary, G. A., and Adams, M. (2009), The Main Sequence of Explosive Active Regions: Discovery and Interpretation, *Astrophys. J.L 700,* L166

Georgoulis, M. K., Rust, D. M. (2007), Quantitative Forecasting of Major Flares, Astrophys. J.L, 661, L109

Georgoulis M. K. (2012), Are Solar Active Regions with Major Flares More Fractal, Multifractal, or Turbulent than Others? Sol. Phys., 276, 161

Hale, G. E., Ellerman, F., Nicholson, S. B., & Joy, A. H. (1919), The Magnetic Polarity of Sun-Spots, *Astrophys. J., 49,* 153

Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. 2006, The Statistical Relationship between the Photospheric Magnetic Parameters and the Flare Productivity of Active Regions, ApJ, 644, 1273

Komm, R., Howe, R., Hill, F., González Hernández, I., & Toner, C. 2005, Kinetic Helicity Density in Solar Subsurface Layers and Flare Activity of Active Regions, ApJ, 630, 1184

Leka, K. D., & Barnes, G. 2003a, Photospheric Magnetic Field Properties of Flaring versus Flare-Quiet Active Regions. I. Data, General Approach, and Sample Results, ApJ, 595, 1277

Leka, K. D., & Barnes, G. 2003b, Photospheric Magnetic Field Properties of Flaring versus Flare-Quiet Active Regions. II. Discriminant Analysis, ApJ, 595, 1296

Leka, K. D., & Barnes, G. 2007, Photospheric Magnetic Field Properties of Flaring versus Flare-Quiet Active Regions. IV. A Statistically Significant Sample, ApJ, 656, 1173

Mason, J. P., & Hoeksema, J. T. 2010, Testing Automated Solar Flare Forecasting with 13 years of Michelson Doppler Imager Magnetograms, ApJ, 723, 634

McAteer, R. T. J., Gallagher, P. T., & Ireland, J. 2005, Statistics of Active Region Complexity: A Large-Scale Fractal Dimension Survey, ApJ, 631, 628

McIntosh, P. S. (1990), The Classification of Sunspot Groups, *Sol Phys. 125*, 251

Reinard, A. A., Henthorn, J., Komm, R., & Hill, F. 2010, Evidence That Temporal Changes in Solar Subsurface Helicity Precede Active Region Flaring, ApJ, 710, L121

Sachs, L. (1978), Applied Statistics: A Handbook of Techniques (Berlin: Springer-Verlag).

Scherrer, P. H., et al. (1995), The Solar Oscillations Investigation Michelson Doppler Imager, Sol. Phys., 162 , 129

Schrijver, C. J. 2007, A Characteristic Magnetic Field Pattern Associated with All Major Flares and Its use in Flare Forecasting, ApJL, 655, L117

Shao, Jun (1993), Linear Model Selection by Cross-Validation, *J. of the American Statistical Association*, Vol. 88, No. 422, 486

Welsch, B.T., Li, Y., Schuck, P.W., & Fisher, G.H. 2009, What is the Relationship between Photospheric Flow Fields and Solar Flares? ApJ, 705, 821

Wheatland, M. S. 2004, A Bayesian Approach to Solar Flare Prediction, ApJ, 609, 1134

Wheatland, M. S. 2005, A Statistical Solar Flare Forecast Method, Space Weather, 3, S07003

Woodcock, F, (1976), The Evaluation of Yes/No Forecasts for Scientific and Administrative Purposes, *MWRv*, 104, 1209

Zirin, H., (1988), Astrophysics of the Sun Cambridge University Press Cambridge, 316