The Astrophysical Journal, 723:1006–1018, 2010 November 10 doi:10.1088/0004-637X/723/2/1006

STATISTICAL PROPERTIES OF SOLAR ACTIVE REGIONS OBTAINED FROM AN AUTOMATIC DETECTION SYSTEM AND THE COMPUTATIONAL BIASES

Jie Zhang[1], Yuming Wang[2], and Yang Liu[3]
[1] Department of Computational and Data Sciences, George Mason University, 4400 University Dr., MSN 6A2, Fairfax, VA 22030, USA; jzhang7@gmu.edu
[2] School of Earth and Space Sciences, University of Science and Technology of China, 96 Jinzhai Road, Heifei, Anhui 230026, China
[3] W. W. Hansen Experimental Physical Laboratory, Stanford University, Stanford, CA 94305, USA
Received 2010 June 29; accepted 2010 September 1; published 2010 October 18

ABSTRACT

We have developed a computational software system to automate the process of identifying solar active regions (ARs) and quantifying their physical properties based on high-resolution synoptic magnetograms constructed from Michelson Doppler Imager (MDI; on board the *SOHO* spacecraft) images from 1996 to 2008. The system, based on morphological analysis and intensity thresholding, has four functional modules: (1) intensity segmentation to obtain kernel pixels, (2) a morphological opening operation to erase small kernels, which effectively remove ephemeral regions and magnetic fragments in decayed ARs, (3) region growing to extend kernels to full AR size, and (4) the morphological closing operation to merge/group regions with a small spatial gap. We calculate the basic physical parameters of the 1730 ARs identified by the auto system. The mean and maximum magnetic flux of individual ARs are 1.67×10^{22} Mx and 1.97×10^{23} Mx, while that per Carrington rotation are 1.83×10^{23} Mx and 6.96×10^{23} Mx, respectively. The frequency distributions of ARs with respect to both area size and magnetic flux follow a log-normal function. However, when we decrease the detection thresholds and thus increase the number of detected ARs, the frequency distribution largely follows a power-law function. We also find that the equatorward drifting motion of the AR bands with solar cycle can be described by a linear function superposed with intermittent reverse driftings. The average drifting speed over one solar cycle is $1.^{\circ}83 \pm 0.^{\circ}04$ yr^{-1} or 0.708 ± 0.015 m s^{-1}.

Key words: Sun: general – Sun: magnetic topology – Sun: surface magnetism

Online-only material: color figures

1. INTRODUCTION

An active region (AR) on the Sun is known as an extended area threaded with strong magnetic fields across the surface. Because of vast free energy stored in these magnetic fields, ARs are the major source of various solar activities, including flares and coronal mass ejections (CMEs), which may further cause severe space weather that adversely affects critical technological systems on the Earth. ARs are historically observed as dark sunspots (due to magnetic cooling effect) in white light images, dating back more than 350 years ago (see Wolf 1861 and Hathaway 2010 for a recent review). Sunspot numbers on the Sun exhibit apparent cyclic behavior with an average period of about 11 years (McKinnon & Waldmeier 1987; Hathaway et al. 1999). The behavior of sunspot positions follows the "Spörer's Law of Zones" as illustrated by the well-known "Butterfly Diagram" (Maunder 1904), that is, sunspots reside in two bands on either side of the equator, and as the cycle progresses, the latitude of sunspot bands expands but slowly drifts toward the equator (Li et al. 2001; Hathaway et al. 2003). The magnetic nature of sunspots was first studied by Hale et al. (1919), who discovered the famous Hale's Polarity Laws. In this paper, we revisit these basic properties of solar ARs and sunspots using modern observational data and the state-of-the-art computational technology for detection and characterization.

In the past, long-term synoptic catalogs of ARs/sunspots were generated through day-to-day manual drawing and later computer-aided inspection of solar images by human operators. There are so far two major data catalogs that have probably served as the foundation of solar physics. One is the International Sunspot Numbers obtained daily since 1849, initiated by Wolf (1861; also called Wolf number or Zürich number) and since 1981 provided by the Royal Observatory of Belgium Solar Influences Data Analysis Center (SIDC). The other one is the National Oceanic and Atmospheric Administration (NOAA) AR catalog produced by the US Air Force and NOAA Space Weather Prediction Center (SWPC) based on images from the Solar Optical Observing Network (SOON) sites since 1977. Each AR in the NOAA catalog is assigned a unique identification number, dubbed as NOAA AR number. The catalog reports ARs' heliographic location, longitudinal extent, sunspot area, and a three-letter classification of sunspots (the so-called modified Zürich classification, or McIntosh classification; McIntosh 1990). Further, the catalog reports the magnetic type of the AR, the so-called alpha–beta–gamma–delta system, based on the coarse magnetic morphology of sunspots. The NOAA AR catalog has been extensively used, in particular, NOAA AR identification numbers have provided a simple but unambiguous cross-reference to specific ARs of the Sun; to date about 11,000 ARs have been reported. Nevertheless, the NOAA AR catalog is a rather basic one, which does not provide quantitative characterization of magnetic properties of identified ARs. Statistical studies of AR magnetic properties were earlier carried out by Howard (1989) based on "coarse array" magnetograms of the Mount Wilson Observatory (MWO) and by Harvey & Zwaan (1993) based on National Solar Observatory (NSO) Kitt Peak (KP) full disk magnetograms.

In recent years, there have been considerable efforts in automating the process of AR identification and characterization (Turmon et al. 2002; Zharkova et al. 2005; McAteer et al. 2005; Colak & Qahwaji 2008), based on magnetograms and white light images obtained since 1996 by the Michelson Doppler Imager (MDI) instrument (Scherrer et al. 1995) on board the *Solar and Heliospheric Observatory* (*SOHO*) spacecraft. The high quality

of space-based data, in combination with modern computational methods in image processing and pattern recognition, enables automated feature and event detection (Aschwanden 2010). Algorithms of automated detection have also been developed and applied for other solar features/events including flares (Qu et al. 2004), filaments/prominences (Qu et al. 2005; Wang et al. 2010), CMEs (Robbrecht & Berghmans 2004; Olmedo et al. 2008; Boursier et al. 2009), and many others (Aschwanden 2010). It has been recognized that the automated detection has become a necessary research tool, especially considering the inception of data at a rate of several terabyte per day from Solar Dynamic Observatory (launched in 2010 February). Without automated feature detection and characterization, bulk volume of solar data obtained by modern instruments will not be inspected by operators and researchers. As a result, rich scientific information residing in these data would not be explored.

A catalog of ARs build upon automated detection and characterization, when compared with the classical NOAA AR catalog, would have obvious advantages in addressing scientific problems. The finding of ARs would be objective and consistent, free from the subjectivity of human inspection. The consistency of data is important for statistical study and modeling long-term evolution of the Sun, i.e., solar dynamo models. For instance, a number of authors have reported different frequency distribution functions of ARs with respect to area size. Tang et al. (1984) reported an exponential distribution of AR area size based on MWO magnetograms. However, based on NSO/KP magnetograms, Harvey & Zwaan (1993) approximated the AR size distribution to be a polynomial function. Based on MWO white light images, Bogdan et al. (1988) found a log-normal distribution of sunspot umbra area size. On the other hand, using MDI white light images and auto-detection method, Zharkov et al. (2005) found that the distribution of sunspot area size is exponential. This kind of controversy also extends to smaller magnetic features, i.e., ephemeral regions and quiet-Sun network features, being either exponential (Schrijver et al. 1997; Hagenaar et al. 2003) or power law (Parnell et al. 2009). Using a combination of *SOHO*/MDI and *Hinode*/Solar Optical Telescope magnetogram data, Parnell et al. (2009) found a power-law distribution of solar magnetic features over more than five decades in flux, expanding from large ARs to small bipolar regions in the quiet Sun. The differences of the distributions presented by these authors are probably caused by different methods employed to identify and characterize ARs. Therefore, it is important to design an appropriate computational method to find ARs, with a full investigation of the possible computational biases of the method used.

Automated detection also facilitates the characterization of ARs, that is, to extract physical properties of an AR in a quantitative way. In computational means, the obtained physical parameters are called the attributes of the AR entity. Since an automated detection defines an AR in pixel level and registers the region as a set of pixels in the input digital image, the calculation of any attribute of the identified region is rather straightforward. Many AR attributes have been proposed in recent years, in order to predict the probability of an AR in producing major flares and/or CMEs. These attributes include the length and gradient magnitude of strong gradient polarity inversion line (PIL; Falconer et al. 2002; Jing et al. 2006; Wang & Zhang 2008), magnetic energy dissipation field (Abramenko et al. 2003), AR fractal dimension (McAteer et al. 2005), effective connected magnetic field (Georgoulis & Rust 2007), unsigned magnetic flux surrounding strong gradient PIL, the so-called *R*-value (Schrijver 2007), and multi-resolution magnetic gradient (Ireland et al. 2008). In particular, based on vector magnetic data from the University of Hawaii Imaging Vector Magnetograph, Leka & Barnes (2003; and in a series of following papers) constructed a large number of AR quantities, including vertical current, current helicity, and magnetic shear angles, to predict AR flare productivity. They found that combinations of only a few familiar variables encompass the majority of the predictive power available, and further concluded that the state of photospheric magnetic field at any given time has limited bearing on whether that region will be flare productive (Leka & Barnes 2007; Barnes & Leka 2008). Further, automated identification of ARs is an essential tool for providing near-real-time prediction of flares and CMEs (Colak & Qahwaji 2009). Flare/CME prediction is at the core of space weather research, one major thrust of solar physics in the past decade.

In this paper, we present an automated AR detection software system and show several preliminary results from the AR catalog generated by this system. The algorithm is based on morphological analysis and intensity thresholding, and is implemented in Interactive Data Language (IDL). The software system is tested and applied to the high-resolution Carrington rotation synoptic magnetogram charts constructed from MDI observations from 1996 to 2008 inclusive. The system is checked and validated against the NOAA AR catalog. In particular, we have carefully investigated computational biases introduced by the selection of controlling or thresholding parameters of the system. A preliminary implementation of the system, but without rigorous validation and bias investigation, has been presented earlier (Wang & Zhang 2008). In Section 2, we present the methodology of the system. Validations and computational biases are discussed in Section 3. In Section 4, we present our characterization of ARs with a set of geometric and flux parameters, and the statistical measures of these parameters. With these parameters consistently obtained through one solar cycle, we further study the solar cycle variation of ARs, AR frequency distributions, and AR-band drifting motion. A summary is provided in Section 5.

2. METHODOLOGY OF THE AUTOMATED AR IDENTIFICATION SYSTEM

Based on the definition that an AR is an extended area of relatively strong magnetic fields, the methodology of identification necessarily involves morphological analysis and intensity thresholding. As illustrated in Figure 1, there are essentially four steps of processing, each of which corresponds to one particular functional module. The results of the modules are shown in panels (b)–(e), respectively, while the input image is in panel (a) and the final output is in panel (f). The input image is a synoptic chart constructed from the stacking of the central meridian strip of the observed full-disk snapshot magnetogram images over the course of 27+ days, encompassing a full Carrington rotation (CR) of the Sun. The sample chart used in Figure 1 is for Carrington rotation 2000 with the starting day on 2003 February 23 and the ending day on 2003 March 19. Through the interpolation of the snapshot magnetograms over one solar rotation, the resulted synoptic chart is a high-resolution 3600 × 1080 pixel map. The *X*-axis is linear in the Carrington longitude (0.1 degree per pixel), while the *Y*-axis is linear in the sine latitude.

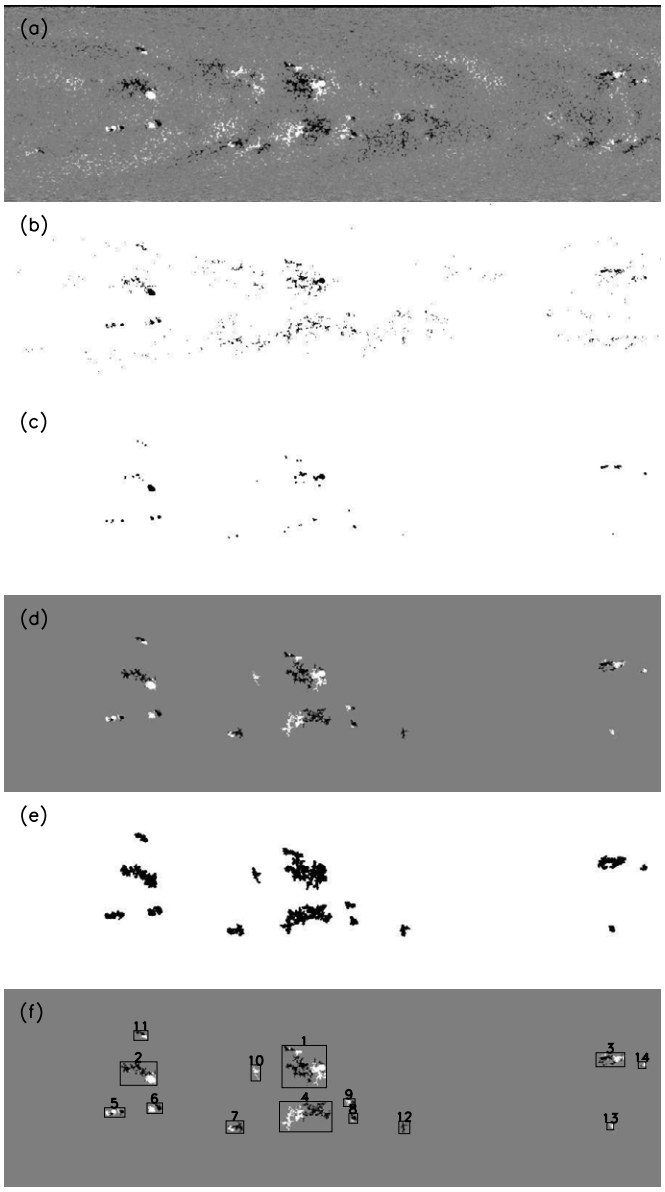

Figure 1. Illustration of the processing modules of the automated AR detection system. (a) Input image. The MDI synoptic map of Carrington rotation 2000 is used here as an example. (b) Module 1: kernel pixels after intensity thresholding segmentation. (c) Module 2: effective AR kernels after the morphological opening operation. (d) Module 3: recover full region size using region growing. (e) Module 4: grouping neighboring regions using the morphological closing operation. (f) Output image: extracted ARs indicated by rectangular boxes and labeled by numbers; each AR is registered as a set of connected pixels.

Compared with a snapshot magnetogram, a synoptic chart has the advantage of being free of projection effect along the longitude, but at the cost of losing temporal resolution. The projection effect along the latitude remains. The projection correction to the field strength has been made during the construction, assuming that the MDI makes line-of-sight measurements of a radial magnetic field. The correction to the pixel area has also been made in our calculation. The underestimation of the flux density in MDI images (Berger & Lites 2003; Tran et al. 2005) has been corrected using a scaling factor derived by Tran et al. (2005), with a combination of a new line-profile saturation factor for the Fe I observation line at 5250 Å (Ulrich et al. 2009). We also point out that the saturation effect due to on-board processing, that the field strength above ∼3400 G appears as a lower field value, has not been corrected.

An automated detection usually requires pre-processing of input images, in order to remove certain artificial defects (e.g., image jitter, corrupted data blocks, cosmic rays, etc.) to reduce the noise and ensure the uniformity of the data. However, we find that the standard high-resolution MDI synoptic charts provided by MDI team satisfy the pre-processing requirement, and therefore no additional pre-processing is applied other than those already made.

The first functional module of the detection system is to apply the intensity segmentation method to isolate kernel pixels. As shown in Figure 1(b), the resulting image is a binary image: kernel pixels forming the foreground in black dots or patches, while all other pixels are set as white background. The intensity thresholding level of the segmentation (I_K) is chosen to be 250 G by trial and error (full justification of the threshold will be given in the following section). The segmentation effectively removes most of the pixels with weak magnetic fields in quiet-Sun regions and polar regions. However, there apparently remain many kernel pixels which did not originate from any AR. These non-AR kernel pixels are mainly from decayed ARs, ephemeral regions, and large flux concentrations in quiet-Sun networks.

The second functional module is to remove these non-AR kernel pixels. Using the fact that AR kernels are more extensive in size, the method is to apply the morphological opening operation on the binary image obtained from the first step. The morphological opening operation is a standard image processing method to remove small geometric objects from the foreground and place them in the background. The effectiveness is controlled by a threshold structural size (S_O), which is chosen to be 10 Mm to obtain the result shown in panel (c). The morphological opening operation consists of two steps of image processing. The first step is the erosion operation (setting the erosional structural size to be S_O) that effectively removes small objects and at the same time shrinks larger objects. The second step is called the dilation operation, which dilates or enlarges the remaining kernel structures by a size of S_O, thus effectively recovering the original size of large kernels. The overall result is that large objects in the input image are almost unchanged, but small objects are permanently removed. The morphological opening operation is essentially a segmentation or filtering based on objects' geometric size but not the intensity. The combination of Module 1 and 2 results in isolated AR kernels.

The third functional module is a standard region-growing operation to recover the full size of an AR from the kernel pixels obtained in Module 2. This is an operation based on intensity and morphology. But the controlling parameter is the intensity threshold, and we choose the threshold to be 50 G (I_A). Starting from kernel pixels that act as seeds, all pixels whose intensities are larger than I_A and that are connected to the seeds are recovered, forming a new gray-scale intensity image shown in panel (d). We find that the threshold of 50 G is a reasonable number for MDI images. This threshold is the same as the one used by McAteer et al. (2005). The ARs recovered (in panel (d)) show good consistency with observations.

The fourth and last functional module is the morphological closing operation, in order to merge neighboring ARs that are very close to each other. Considering the likely interconnection of magnetic fields in the corona from these neighboring regions that might collectively determine their solar activities, we should treat these regions as a single entity. The morphological closing operation is the opposite of the opening operation, that is, to

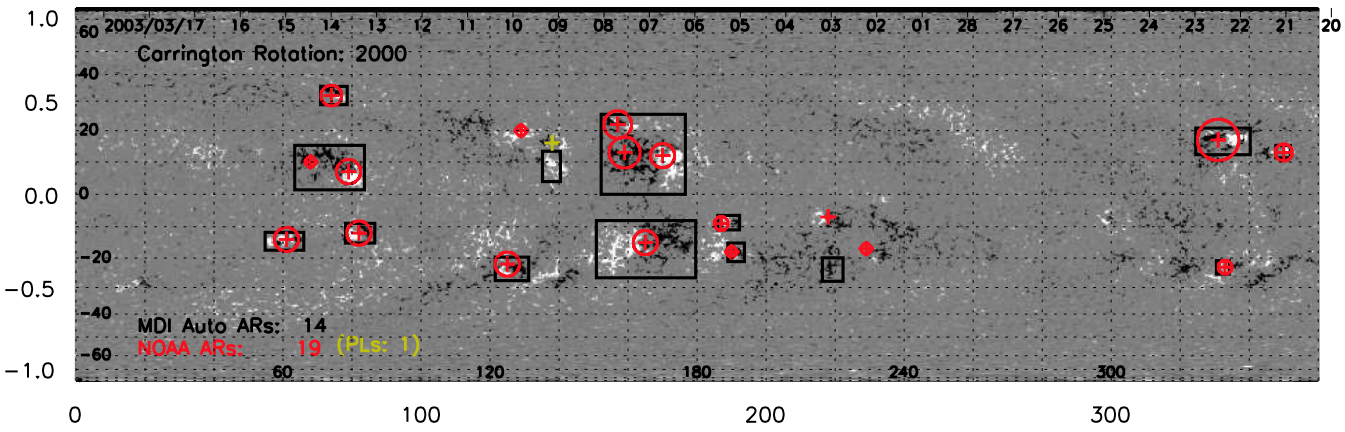

Figure 2. ARs identified by the automated method (black boxes) vs. that from the NOAA catalog made by human inspectors overlaid on the gray-scale synoptic magnetogram map of CR 2000. The red circles (and plus symbols at the center) indicate the relative size (and centroids) of the NOAA ARs; the only non-spot plage region is denoted by the yellow symbol.

(A color version of this figure is available in the online journal.)

apply the dilation first and the erosion second. The structural size of this operation (S_C) is chosen to be 10 Mm, which means that any gap smaller than S_C will be repaired and the two regions merged into a single AR. After this operation, any AR must be separated from other ARs by at least 10 Mm in space. The final result is shown in panel (f). Identified ARs are indicated by black rectangular boxes labeled with numbers in the order of decreasing size.

Note that the method presented above is general enough to find any features or objects based on intensity and size, so that it can be applied to many other types of images, including snapshot MDI magnetogram images for finding ARs and MDI white light images for finding sunspots. Indeed, we are in the process of developing an automated tracking module using time series of snapshot magnetogram images to track the evolution of an AR as it traverses across the front-side disk of the Sun. The result of such tracking will be presented in a future paper.

3. VALIDATION AND COMPUTATIONAL BIASES

3.1. Validation Metrics

The four-step identification method discussed above involves a set of four and only four controlling or thresholding parameters. We have used the set of $I_K = 250$ G, $S_O = 10$ Mm, $I_A = 50$ G, and $S_C = 10$ Mm to illustrate the processes as shown in Figure 1. Apparently, a different set of controlling parameters would produce a different detection result, in terms of not only the number of ARs, but also the size of the regions detected. While an automated method is free from the subjectivity of human operators in identifying ARs, there is certainly a computational bias associated with the selection of controlling parameters. Here we vigorously investigate such computational bias.

In this paper, we use the standard contingency table to validate our auto-detection system, assuming that the NOAA catalog provides the "ground truth." A contingency table contains four parameters: (1) the number of true positives (TPs) or hits, (2) the number of false positives (FPs), or false alarmings, also called Type-I errors, (3) the number of false negatives (FNs), or missing detections, also called Type-II errors, and (4) the number of true negatives (TNs). Using CR 2000 as an example (Figure 2), the automated method identifies 14 ARs ($N_{\rm AUTO}$, indicated by the black rectangular boxes), while the NOAA catalog reports 19 ARs ($N_{\rm NOAA}$, indicated by the red circular symbols, while the sole yellow symbol is for a non-spot AR). For this rotation, the number of TP prediction is 15 ($N_{\rm TP}$), meaning that 15 NOAA ARs have their reported center locations within the boxes of automated ARs. Correspondingly, the number of FNs is 4 ($N_{\rm FN}$), the NOAA ARs that are not caught by the auto method. The number of FPs is two ($N_{\rm FP}$), meaning that two auto ARs do not contain any NOAA ARs. While the above three parameters are well defined, the number of TNs ($N_{\rm TN}$) cannot be determined by the system, since there is no negative event identified in the context of AR finding. A negative AR in a solar image would correspond to any surface area outside positively identified ARs, and thus is not a constrained entity. Nevertheless, for the convenience of this discussion, we arbitrarily assume that $N_{\rm TN}$ equals the number of observed positive events, or $N_{\rm NOAA}$; thus, it is treated as a constant independent of the selection of controlling parameters.

Based on the contingency table, several validation metrics can be constructed. One popular metric is the true positive rate $R_{\rm TP}$, which measures the success of finding NOAA ARs, defined as

$$R_{\rm TP} = \frac{N_{\rm TP}}{N_{\rm TP} + N_{\rm FN}} = \frac{N_{\rm TP}}{N_{\rm NOAA}}. \quad (1)$$

This parameter also denotes the rate of missing events, which equals $1 - R_{\rm TP}$. Another popular metric is the false positive rate, which shows the rate of overdetection, or false alarming of the auto system. It is defined as

$$R_{\rm FP} = \frac{N_{\rm FP}}{N_{\rm FP} + N_{\rm TN}} = \frac{N_{\rm FP}}{N_{\rm FP} + N_{\rm NOAA}}. \quad (2)$$

Therefore, for the set of controlling parameters chosen above and for CR 2000, our auto-detection system yields a true positive rate of 78.9% and a false positive rate of 9.5%. A desirable system is to maximize the true positive rate and at the same time minimize the false negative rate.

We further introduce one more metric, the rate of compound AR (CAR; $R_{\rm CAR}$), to address the issue of multiple-to-one mapping between NOAA ARs and auto ARs. For example, as shown in Figure 2, the largest AR (region number "1" in Figure 1(f)) contains three individual NOAA ARs. This is caused by the fact that a magnetic region seen in the magnetogram is much more extensive in size than the corresponding sunspot seen in white light (the size ratio is about 15, as discussed in the following

section). As a result, while ARs appear as discrete sunspot groups in white light images, they may appear morphologically connected in magnetogram images. In particular, when the magnetic fields expand out into the corona, they may be interconnected and collectively determine solar activities. Therefore, for both computational vigor and physical justification, the multiple-to-one mapping is necessary in our system. In CR 2000, there are two ARs having such mapping. We may call these regions as CARs. Therefore, we introduce the rate of CAR as

$$R_{\rm CAR} = \frac{N_{\rm CAR}}{N_{\rm AUTO}}. \quad (3)$$

Thus, the rate is 14.3% for the example discussed.

We here present the variation of these metrical parameters with the solar cycle from CR 1909 (starting on 1996 May 5) to CR 2077 (ending on 2008 December 17) for the set of controlling parameters chosen above (Figure 3). The period after 2008 December is largely within the extended solar minimum, during which there is almost no AR. The thirteen-year-long period from 1996 to 2008 encompasses the entire 23rd solar cycle starting from its minimum and extending into the beginning minimum of the 24th solar cycle. Apparently, the AR solar cycle variation from our auto detection closely follows that reported by NOAA (Figure 3, top panel). Note that there are several missing data points from the auto detection (CRs 1937, 1938, 1939, and 1940), which are caused by the malfunction of the *SOHO* spacecraft in 1998. Further, the following CRs, 1941, 1944, 1945, 1956, 2011, and 2015, are not included in the plot because of incomplete MDI data in these rotations. We simply replace these missing data points with the value of previous effective data points. Further, in order to better view the solar cycle trend, we have applied a six-point running average on all the profiles shown in Figure 3.

Throughout the solar cycle, the true positive rate $R_{\rm TP}$ varies between ~40% and 90%, with an average value at 73.8%; the average is weighted by the number of auto ARs per Carrington rotation. On the other hand, the false positive rate $R_{\rm FP}$ varies between ~0% and ~25%, with an average value at 15.3% (also weighted by the number of ARs). The rate of compound ARs varies between zero and ~20% with an average value at 13.0% (also weighted by the number of ARs). Apparently, this rate is very low during the solar minima when ARs are sparse on the Sun, and becomes relatively large during the solar maximum when a large number of ARs are simultaneously present on the Sun.

3.2. Computational Biases

With the validation metrics defined above, we now consider the computational bias of our auto system, when different sets of controlling parameters are used (Figure 4). Each metric parameter shown in the figure is the average value obtained over the whole solar cycle from 1996 to 2008. In the first column, we show the variation of the metric parameters by varying the AR kernel pixel intensity threshold I_K from 100 G to 400 G, while the other controlling parameters are fixed (the size of the morphological opening operation $S_O = 10$ Mm, the AR pixel intensity threshold $I_A = 50$ G, and the size of the morphological closing operation $S_C = 10$ Mm). The four panels from top to bottom show the following parameters, respectively: (1) the ratio ($R_{\rm AR}$) between the auto AR number ($N_{\rm AUTO}$) and the NOAA AR number ($N_{\rm NOAA}$), (2) the true positive rate ($R_{\rm TP}$), (3) the false negative rate ($R_{\rm FN}$), and (4) the rate of compound ARs ($R_{\rm CAR}$).

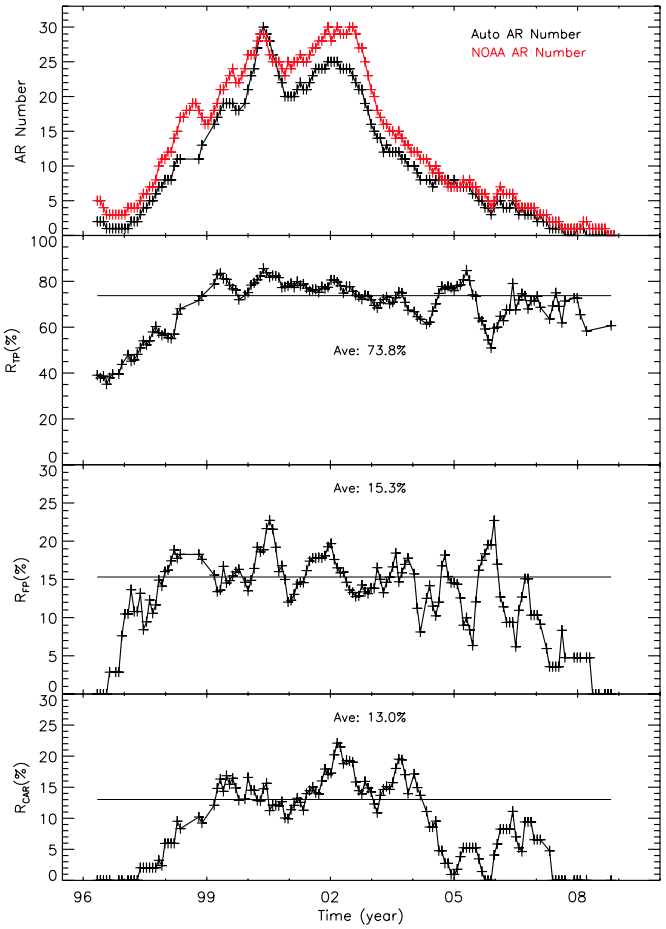

Figure 3. Validation metric parameters of the automated detection system against the NOAA AR catalog. The parameters are obtained for every Carrington rotation from 1996 to 2008. The four panels from top to bottom are for auto-detected AR numbers (black line, NOAA AR numbers in the red line), true positive rate, false positive rate (or "missing"), and the rate of compound ARs, respectively. The four controlling parameters for this instance of calculation are 250 G, 10 Mm, 50 G, and 10 Mm, respectively.

(A color version of this figure is available in the online journal.)

As seen in the figure, when I_K increases from 100 G to 400 G, the number of auto-detected ARs decreases from 4320 to 1060 ($N_{\rm AUTO}$), corresponding to $R_{\rm AR}$ from ~2.1 to 0.5. Note that during the period from 1996 May 5 to 2008 December 17, there were in total 3048 ARs reported in the NOAA catalog. Among these regions, 2286 ARs had crossed the central meridian, while the others either disappeared before or emerged after the central meridian. Since we use only Carrington synoptic maps in this study, the non-central-meridian-crossing ARs should not be used in the comparison. Further, excluding those CRs with no or incomplete MDI observations, the number of NOAA ARs used in our comparative study stands at 2085 ($N_{\rm NOAA}$). Apparently, the number of ARs auto-detected is rather sensitive to the kernel pixel intensity threshold. As the threshold decreases, regions with weaker magnetic fields will be included. These regions may not correspond to any NOAA ARs. When $I_K = 100$ G, the resulted $N_{\rm AUTO} = 4320$, and the false positive rate $R_{\rm FP}$ is as high as 55%, meaning that about half of auto ARs are not co-spaced with any NOAA AR. Nevertheless, the true positive rate $R_{\rm FP}$ reaches almost 90%. In contrast, when the kernel intensity threshold is chosen to be 400 G, almost all small regions including small NOAA ARs are rejected; only large NOAA ARs survive the auto-detection system. As a result, the

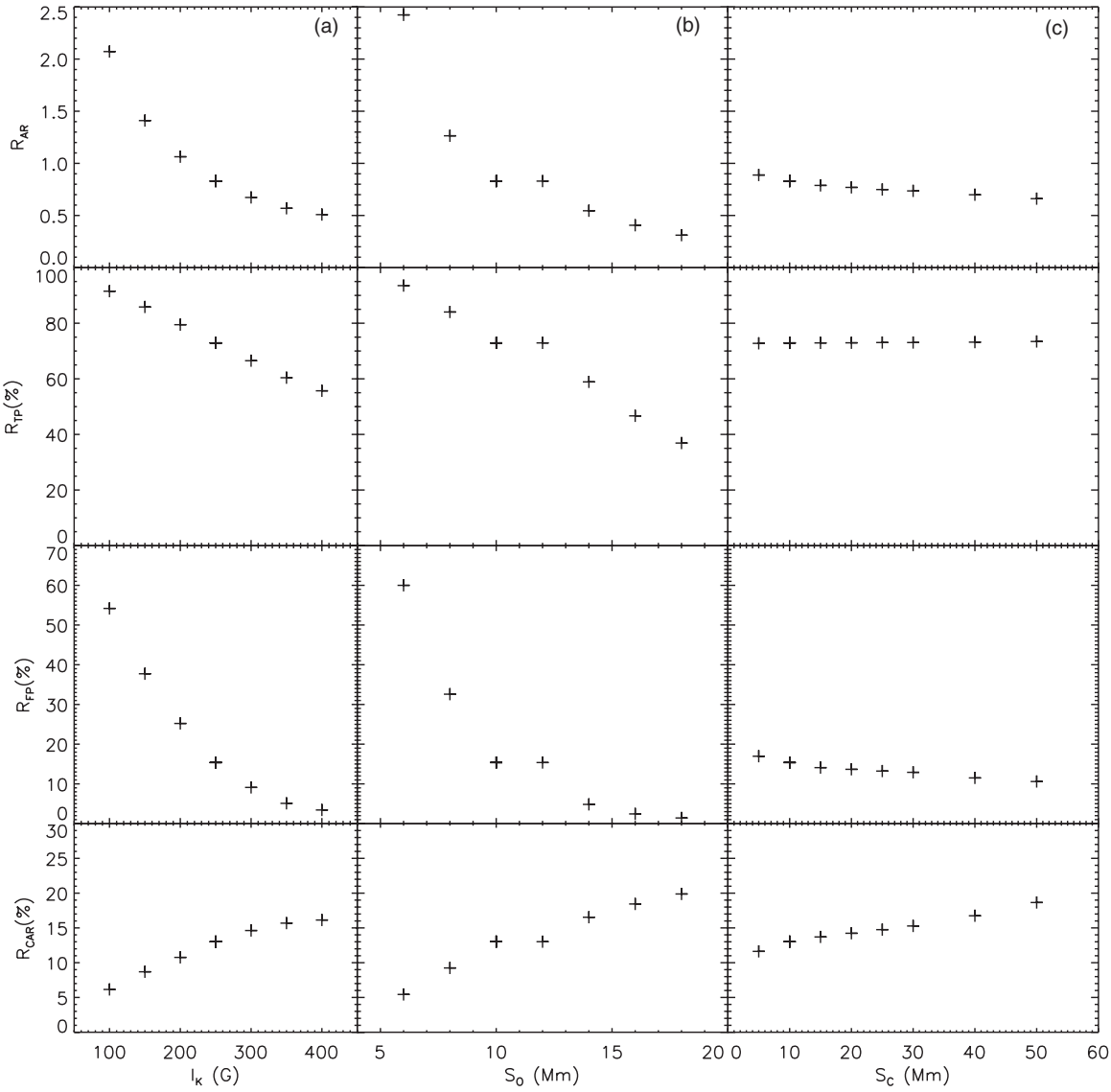

Figure 4. Variation of the metric parameters with respect to the controlling parameters. From top to bottom, the four panels show the ratio between the auto AR number and the NOAA AR number, true positive rate, false positive rate, and compound AR rate, respectively. The three columns from left to right are of varying I_K, varying S_O, and varying S_C, respectively. In all calculations, the minimum AR intensity threshold I_A is set at 50 G.

rate of true positive is 45%, meaning that about 55% of NOAA ARs are missed in the detection. However, the rate of false positive is extremely low (close to zero), meaning that almost every auto AR has corresponding NOAA ARs.

In the middle column of Figure 4, similar variations are shown for controlling parameter S_O, when it increases from 6 Mm to 18 Mm, while the other parameters are fixed ($I_K = 250$ G, $I_A = 50$ G, and $S_C = 10$ Mm). As the morphological opening size increases, the number of ARs auto-detected decreases from 5053 to 649, corresponding to the ratio R_{AR} decreasing from 2.4 to 0.3. In the mean time, the true positive rate decreases from ∼95% to 35%, and the false positive rate decreases from ∼60% to almost zero. It demonstrates that the auto-detection system is sensitive to the size of the morphological opening operation.

As seen in the right column of Figure 4, our detection system only weakly depends on S_C, the structural size of the morphological closing operation. As the size increases from 5 Mm to 50 Mm, the number of ARs detected decreases from 1851 to 1309, corresponding to the AR ratio R_{AR} from 0.89 to 0.66. The true positive rate remains ∼73%, and the false positive rate decreases from ∼17% to 10%.

In all cases we have investigated, we choose the minimal AR pixel intensity threshold I_A at 50 G. This is an arbitrary but reasonable election. It is about three times as large as the standard deviation of the magnetic fields in MDI magnetogram images. Since this parameter is only used on the region-growing operation, it has minimal impact on the number of ARs detected. Nevertheless, it may affect the characterization of an AR, e.g., its area size and total magnetic flux. A lower value would make an AR grow larger, and a larger value would make an AR appear smaller.

Now, the essential question is what controlling parameters we should adopt. There is no simple answer to this. The auto-detected regions which are not in the NOAA AR catalog are also magnetic features with true physical meaning (e.g., ephemeral regions); they are not noisy features. These regions should be included if the purpose is to study any sizable magnetic feature on the Sun. Apparently, the NOAA AR catalog

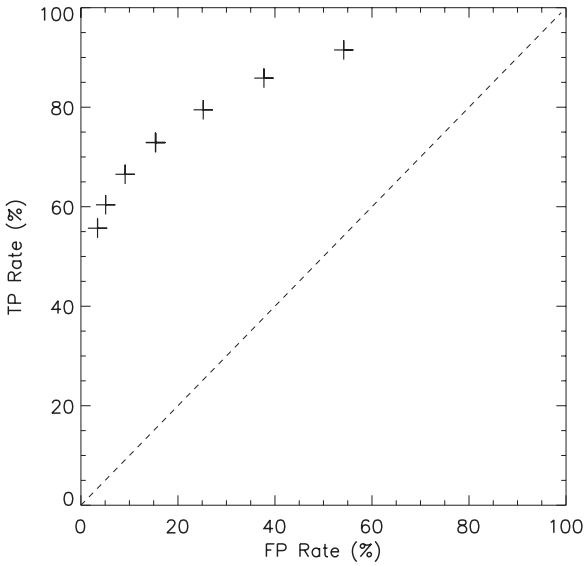

Figure 5. ROC plot of the true positive rate (*Y*-axis) vs. the false positive rate (*X*-axis), varying the kernel pixel intensity threshold I_K from 100 G to 400 G, with an increment of 50 G between neighboring data points. The diagonal broken line denotes the no-discrimination line.

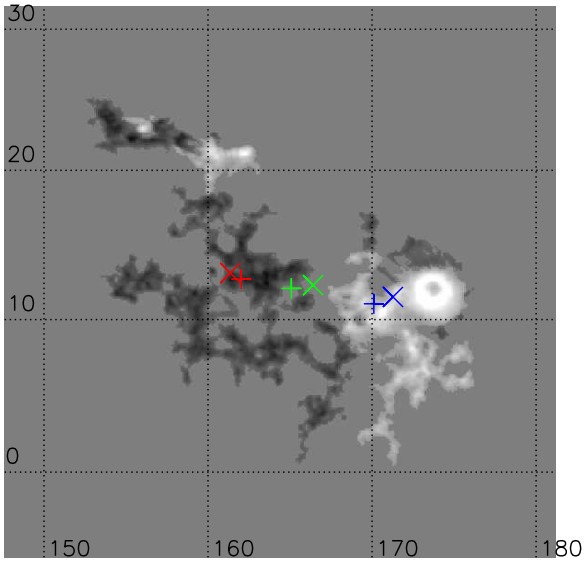

Figure 6. One auto-detected AR in Carrington rotation 2000. The geometric centroids and flux-weighted geometric centroids are indicated by the plus and cross symbols, respectively. The three colors, red, blue and green, are of negative, positive, and unsigned magnetic fluxes, respectively.

(A color version of this figure is available in the online journal.)

contains only large magnetic regions on the Sun, further biased toward those having obvious sunspots in white light. If we assume that the NOAA AR catalog is a "perfect" catalog or the ground truth, then a "perfect" auto-detection system would have the true positive rate at 100% and the false positive rate at 0%.

One popular method to optimize controlling parameters is to analyze the so-called Receiver Operating Characteristic (ROC) plot, or simply, the ROC curve. We show one such plot in Figure 5, which shows the true positive rate (*Y*-axis) versus the false positive rate (*X*-axis) by varying the AR kernel intensity threshold I_K (while fixing other parameters, as for the plots in the right column of Figure 4). The seven data points (indicated by the cross symbols) in the figure correspond to I_K from 100 G to 400 G, with an increment of 50 G for two neighboring points. The optimization is to find the data point that is farthest from the diagonal line (dotted line). The diagonal line is called the no-discrimination line, since a data point on the line would have the same true positive rate as the false positive rate; in other words, the benefit is canceled out by the cost.

Apparently, the seven data points are almost in parallel with the no-discrimination line, with the point in the middle (I_K = 250 G) slightly further from the line. Therefore, from the benefit–cost point of view, there is no strong preference for any of the seven points. The decision is largely based on physical justification, depending on who prefers finding only large ARs (large threshold), or finding all ARs including smaller magnetic features (small threshold) at the cost of significant overdetection. In this study (including AR characterization discussed in the following section), we choose controlling parameters that make this compromise, i.e., intermediately high true positive rate, intermediately low false positive rate, and similar number of auto ARs as that of NOAA ARs. Because of this consideration, the "optimized" set of controlling parameters is found to be (I_K = 250 G, S_O = 10 Mm, I_A = 50 G, and S_C = 10 Mm) and the resulting validation parameters are ($R_{\rm AR} = 0.86$, $R_{\rm TP}$ = 73.8%, $R_{\rm FP}$ = 15.3%, and $R_{\rm CAR}$ = 13.0%).

4. STATISTICAL PROPERTIES OF ARs

The auto-detection system, as presented above, works on two-dimensional digital images on the pixel level. Any identified AR is registered explicitly by a set of morphologically connected pixels in the image. With such a registration, one can further calculate parameters to characterize the physical properties of the AR.

4.1. Basic AR Parameters

For each AR identified by the automated system, we calculate three sets of parameters, which characterize the basic physical properties of the AR: (1) location, (2) geometric size, and (3) magnetic flux. In Figure 6, we show in detail one particular AR as a sample before we discuss the statistical properties of all ARs. This region is the largest AR in CR 2000, denoted as region 1 in Figure 1(f).

The geometric center of the AR (green plus symbol) is located at (165.°1, 12.°0) (the two numbers are the Carrington longitude and latitude, respectively). On the other hand, the flux-weighted geometric center (green cross symbol) is at (166.°4, 12.°2) or 16.0 Mm apart from the plain geometric center. The offset to the right is apparently caused by the strong flux concentration within the leading positive polarity. The geometric center for the leading positive polarity (blue plus symbol) is located at (170.°1, 11.°0), while for the trailing negative polarity (red plus symbol) is (162.°0, 12.°6). Thus, the distance between the leading and trailing polarities, as defined by the geometric centers, is 98.1 Mm. When we define the distance using the flux-weighted geometric centers, the value is as large as 119.2 Mm.

It is also straightforward to calculate the geometric area sizes and the magnetic fluxes threading through the area. The geometric area sizes for the total unsigned, positive, and negative fluxes for this AR are 22489.7 $\rm Mm^2$, 8507.7 $\rm Mm^2$, and 13981.9 $\rm Mm^2$, respectively. The corresponding magnetic fluxes are 6.59×10^{22} Mx, 3.35×10^{22} Mx, and 3.23×10^{22} Mx, respectively. It is interesting to note that, while the ratio between the positive and negative magnetic fluxes is nearly unity (1.04

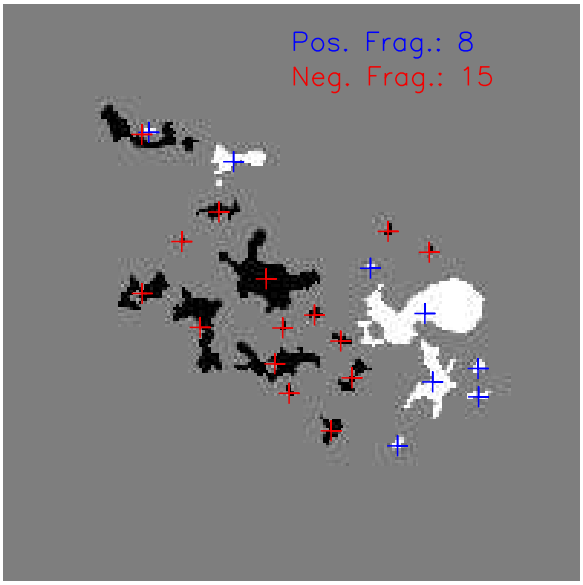

Figure 7. Same AR as in the previous figure but is processed to show the "patchy" magnetic fragments: negative polarity (black) and positive polarity (white). The plus symbols indicate the geometric centroids of the fragments.

(A color version of this figure is available in the online journal.)

to be exact), the ratio of the areas is as large as 1.64. Apparently, the flux of the leading (positive) polarity is concentrated in a relatively smaller area, while that of the trailing polarity disperses more. This kind of leading–trailing asymmetry is a well-known fact of bipolar ARs.

Further, we have applied morphological analysis on individual ARs (Figure 7). The process is almost identical to the one used for the entire synoptic chart as discussed in Section 2, except for the difference of the controlling thresholding parameters. While the two intensity controlling parameters are the same: 250 G for the kernel and 50 G for the nominal AR pixel, the two size parameters are much smaller, both of which are 2 Mm (instead of 10 Mm for AR identification). As shown in Figure 7, we are now able to identify small fragments in the AR. There are 8 fragments of positive polarity (indicated by the white patches and blue symbols at the center of each patch) and 15 fragments of negative polarity (dark patches and red symbols). Therefore, there are in total 23 fragments in this AR. It illustrates that, even in an AR which appears as a single big clump of magnetic flux over the surface of the Sun, the magnetic flux is far from an even distribution in space. The magnetic flux in an AR is fragmented and highly clumpy, concentrated in small patches. This kind of fractal property of ARs has been studied by several authors (Lawrence et al. 1993; Meunier 1999; McAteer et al. 2005; Conlon et al. 2008). In white light images, it is well known that an AR contains multiple sunspots. In a sense, these fragments are analogous to individual sunspots in an AR. To understand the relation between these magnetic fragments and individual sunspots require a study of direct comparison between magnetogram images and white light images, which is beyond the scope of this paper.

4.2. AR Solar Cycle Variation and the Statistical Measures

With the basic parameters obtained for all ARs, we now investigate their global and statistical properties. In Figure 8, we show the variation of AR number, number of fragments, area size, and magnetic flux per Carrington rotation with respect to the phase of solar cycle from 1995 to 2008. In Table 1, we further

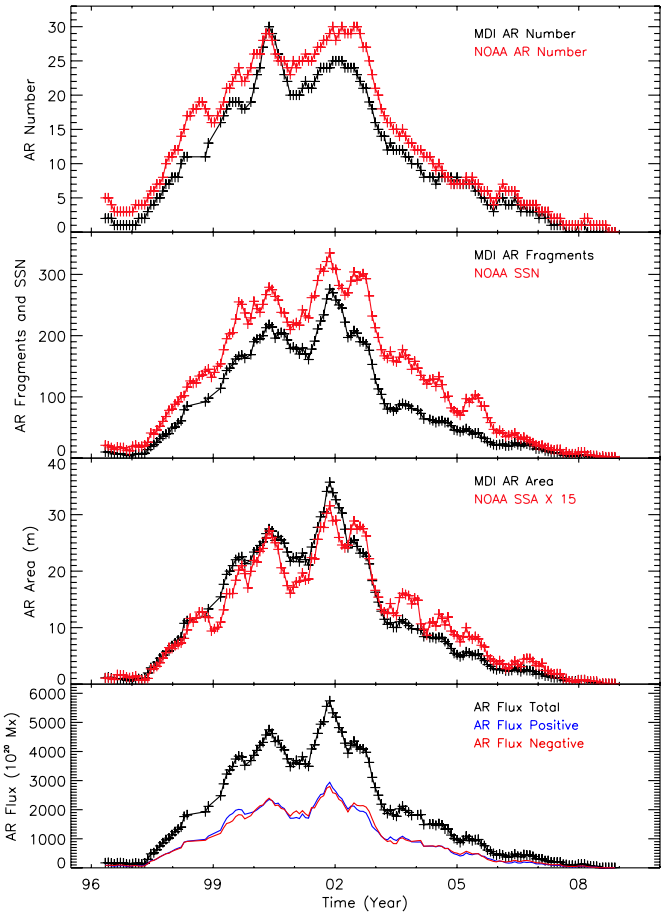

Figure 8. Solar cycle variation from 1996 to 2008 of the per-Carrington-rotation AR parameters. The four panels from top to bottom show (1) the auto-detected AR number (black) and NOAA AR number (red), (2) auto-detected AR magnetic fragments (black) and NOAA sunspot numbers (red), (3) AR area size (black) and NOAA sunspot area size (red), the magnetic area is about 15 times as large as the corresponding sunspot area, and (4) AR unsigned flux (black), positive flux (blue), and negative flux (red).

(A color version of this figure is available in the online journal.)

list the statistical measures of these basic parameters averaged over the solar cycle. Corresponding data from the NOAA AR catalog are also shown as a comparison.

As shown in Figure 8 (top panel), there is apparently a good correlation between the auto (black line) and NOAA (red line) AR numbers, since they both show very similar solar cycle variation. Both numbers indicate that there are double peaks during the solar maximum, one in early 2000 and the other in late 2001; the separation of the two peaks is about one and a half years. In the NOAA AR numbers, the two peaks have almost equal strength, while in the auto AR numbers, the first peak is stronger. There are typically about 30 ARs per Carrington rotation during the solar maximum, while there are only a few or zero during the solar minima at 1996 and 2007/2008. The highest AR number per Carrington rotation is 39 from the NOAA catalog and 37 from our automated method. The AR number averaged over the whole solar cycle is 13.5 from the NOAA catalog and 10.9 from the automated method.

In the second panel of the figure, we show the number of AR fragments per rotation from the automated method (black line) and the number of sunspots (red line) per rotation from the NOAA catalog. There are about 300 sunspots per rotation during the solar maximum, while the number of fragments is

Table 1
Statistical Measures of Solar ARs from 1996 to 2008

Parameter	Method[a]	Mean	Median	Min	Max	STD
AR number	Auto-CR	10.9	8.0	0	37	9.3
	NOAA-CR	13.5	11.0	0	39	10.2
Fragment	Auto-CR	85.3	55.0	0	327	86.3
SSN	NOAA-CR	126.2	101.0	0	413	86.3
Fragment	Auto-AR	7.8	5.0	1	69	8.5
SSN	NOAA-AR	8.0	4.0	0	90	11.0
Area (m)[b]	Auto-CR	10.8	7.0	0	43.3	11.0
	NOAA-CR	0.73	0.53	0	2.95	0.71
Area (μ)[c]	Auto-AR	990.1	569.1	40.1	10994.5	1197.1
	NOAA-AR	43.6	10.0	0	1120.0	89.5
Flux (10^{20} Mx)	Auto-CR	1826.1	1255.5	0	6962.6	1825.4
	Auto-AR	166.8	84.1	8.6	1974.0	219.1

Notes.
[a] Auto: automated method; NOAA: from NOAA manual catalog; CR: on the basis of Carrington rotation; AR: on the basis of individual ARs.
[b] milli, in units of thousandth of the total solar surface area.
[c] micro, in units of millionth of the total solar surface area.

slighter smaller. The highest sunspot number per rotation is 413, while the highest fragment number per rotation is 327. In terms of mean value over the solar cycle, the sunspot number per rotation is 126.2, and the fragment number is 85.3. We believe that counting small fragments within an AR is probably more accurate with the automated method. In the NOAA catalog, the following formula, the so-called Zürich method (e.g., Hathaway 2010), is used to define the sunspot number:

$$R = k(10g + n), \tag{4}$$

where g is the number of sunspot groups, n is the number of individual sunspots, and k is a correction factor for the observer. It is recognized that, when solar images are inspected by human eyes, it is far easier to identify sunspot groups than to identify each individual sunspot. However, the automated method is able to identify individual fragments efficiently and objectively, free of the rather arbitrary assignment of sunspot groups. On the basis of individual ARs, Table 1 shows that an AR has 8.0 sunspots on average, similar to the 7.8 fragments per AR obtained using our automated method. The maximum number of sunspots an AR can contain is 90, while the maximum number of fragments of an AR is 69.

In the third panel of Figure 8, we show the solar cycle profile of the total geometric area of ARs (black line) per Carrington rotation; the unit is in milli, or one-thousandth of the total solar surface (TSS). Magnetic ARs occupy about 30 thousandth or 3% of the total solar surface during the solar maximum. This number is about 15 times larger than the area occupied by the sunspots (red line) in white light as reported by NOAA. Throughout the solar cycle, there is obviously a similar ratio between the AR magnetic area and the sunspot area. The mean area per Carrington rotation over the solar cycle is 10.8 thousandth of the TSS using our automated method, and is only 0.73 thousandth from the NOAA catalog. The maximum value of area per rotation is 43.3 and 2.95 thousandth for the automated and NOAA ones, respectively. All these numbers consistently show that an AR defined by extended magnetic fields in a magnetogram image is about 15 times as large as that defined by the dark region in a corresponding white light solar image.

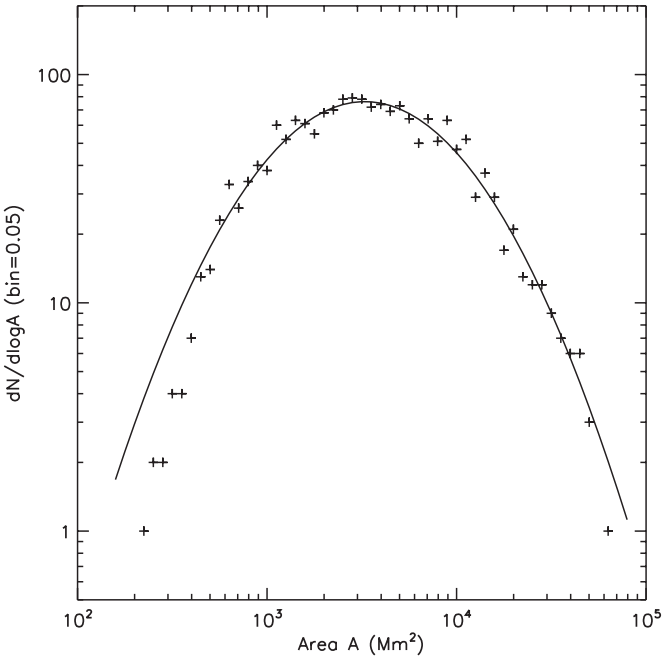

Figure 9. Area size distribution of ARs obtained from the automated system. The data points are well fitted with a log-normal function (black curve). The set of four controlling parameters of this instance of calculation are 250 G, 10 Mm, 50 G, and 10 Mm, respectively.

In the last panel of the figure, we show the solar cycle variation of AR magnetic fluxes derived from the automated method. The three lines represent the total unsigned flux (black), positive flux (blue line), and negative flux (red line) per Carrington rotation, respectively. Note that there is no report on magnetic flux from the NOAA catalog. During the solar maximum, there are about 5×10^{23} Mx of magnetic flux concentrated in ARs per Carrington rotation. The mean magnetic flux per rotation over the solar cycle is 1.83×10^{23} Mx and the maximum magnetic flux is 6.96×10^{23} Mx. For individual ARs, the mean magnetic flux is 1.67×10^{22} Mx and the maximum magnetic flux is 1.97×10^{23} Mx.

It is interesting to point out that, while there are double peaks during the solar maximum with similar strength in the NOAA AR number, the other parameters, including sunspot number, AR area, and flux, show an outstanding peak in late 2001. The strength of this peak is significantly larger than that in early 2000. It has been noted that the dates of solar cycle maxima, when they are determined by difference indexes, i.e., sunspot numbers versus sunspot areas, could be significant different, from a few months to a few years (Hathaway 2010). Further, we note that the peak magnetic flux in late 2001 is mainly caused by the large area size of the emerged ARs, but not the mean intensity of the magnetic field. In terms of the amount of magnetic flux emerged onto the surface of the Sun, it is fair to say that the 23rd solar cycle peaked in late 2001 but not in early 2000.

4.3. AR Frequency Distributions

Statistical frequency distributions of AR sizes are studied in this subsection. There are 1730 ARs identified by the automated system based on MDI high-resolution synoptic maps from 1996 to 2008, using the "optimized" set of controlling parameters. We show the area size distribution and magnetic flux distribution of these ARs in Figures 9 and 10, respectively, plotted in a log–log

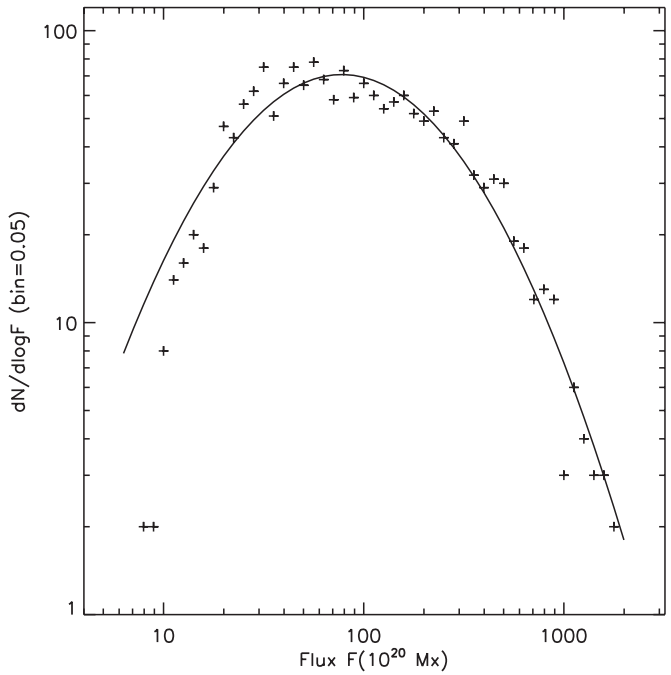

Figure 10. Magnetic flux distribution of ARs obtained from the automated system. The data points are well fitted with a log-normal function (black curve). The set of four controlling parameters of this instance of calculation are 250 G, 10 Mm, 50 G, and 10 Mm, respectively.

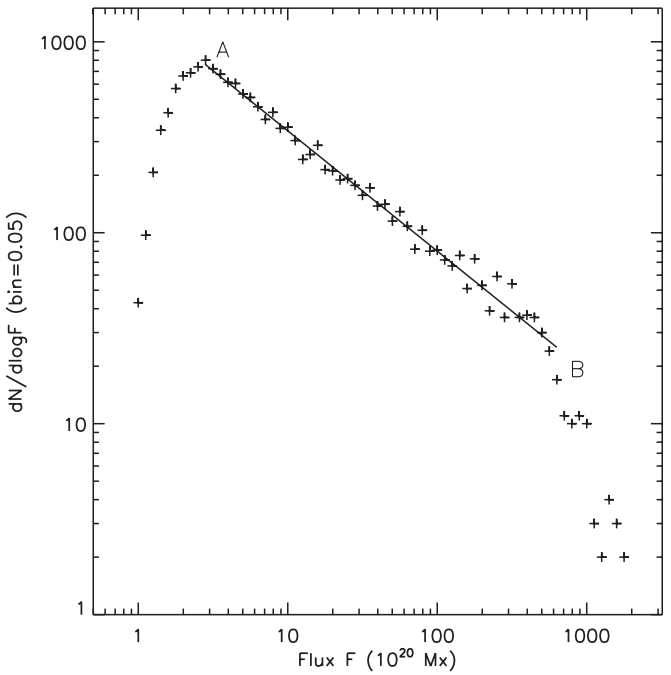

Figure 11. Magnetic flux distribution of magnetic features obtained from the automated system. The set of four controlling parameters of this instance of calculation are (150 G, 6 Mm, 50 G, and 6 Mm), respectively, which are smaller than that used for best reproducing the NOAA AR catalog. Except for the points at the two extreme ends, most data points are well fitted by a straight line, following the power-law function.

scale. The two distributions appear similar and to be Gaussian-like. Therefore, we fit the distribution to a log-normal function. For the area size, the fitted function is

$$\frac{dN(A)}{d\log A} = 76.2 \times \exp\left\{-\frac{(\log A - 3.52)^2}{0.454}\right\}, \quad (5)$$

where A denotes the AR area in units of Mm^2. The bin size of the distribution with respect to $\log A$ is 0.05. The reduced χ^2 test of the goodness of fit is 25.2. For the magnetic flux, the fitted function is

$$\frac{dN(F)}{d\log F} = 70.8 \times \exp\left\{-\frac{(\log F - 1.89)^2}{0.541}\right\}, \quad (6)$$

where F denotes the unsigned AR magnetic flux in units of 10^{20} Mx. The bin size of the distribution with respect to $\log F$ is 0.05. The reduced χ^2 test of the goodness of fit is 57.2.

The log-normal distribution is consistent with that found by Bogdan et al. (1988). However, it is different from the functions suggested by other workers. The exponential distribution of ARs was found by Tang et al. (1984), Schrijver et al. (1997), and Zharkov et al. (2005). The polynomial function was used by Harvey & Zwaan (1993) to fit their data. The power-law distribution was found by Parnell et al. (2009). The functional form of AR distribution may reveal the physical mechanism that generates these ARs. Bogdan et al. (1988) argued that the log-normal distribution is a result of magnetic fragmentation in the convection zone. On the other hand, Schrijver et al. (1997) demonstrated that frequent fragmentation and collision (or merging) of magnetic features lead to the exponential distribution, favoring the cause of near-surface dynamo. Our data indicate that large ARs tend to be caused dominantly by the fragmentation process, thus favoring the global dynamo model occurring through the convection zone (Bogdan et al. 1988).

Nevertheless, our fitting has a large χ^2 value, indicating that the log-normal function is a rather poor fitting to the observed data. Careful inspection of Figures 9 and 10 reveals that the fitting is better in the higher end of the distribution than in the lower end. Toward the lower end, the frequency distribution drops much faster than the normal distribution. We believe that the fast drop is caused by the computational bias. For this instance of automatic detection, the set of controlling parameters is chosen to be (250 G, 10 Mm, 50 G, and 10 Mm), which favors the detection of large magnetic features, while selectively remove features of small size and weak intensity.

To further illustrate the influence of computational bias, we have made one instance of automatic detection with significantly reduced thresholds. We choose the set of controlling parameters to be (150 G, 6 Mm, 50 G, and 6 Mm), thus favoring the detection of weaker and smaller magnetic features. The smaller structural size for the morphological opening operation makes the system not only detect small non-AR features, but also tends to fragment an AR into multiple pieces. The smaller structural size for the morphological closing operation also reduces the possibility of grouping neighboring magnetic features. As a result, the automated system now finds 14,431 magnetic features, about 8.3 times as many as the number of ARs found at the instance of (250 G, 10 Mm, 50 G, and 10 Mm). Most of the detected magnetic features are not nominal ARs; they could include fragments of decayed ARs, ephemeral regions, and even large network features in the quiet Sun.

In Figure 11, the frequency distribution of these features with respect to magnetic flux is shown. Apparently, the system can detect magnetic features with flux as small as 1.0×10^{20} Mx, which is about eight times smaller than that from the instance for ARs. Interestingly, with this instance of detection, the AR distribution now becomes largely a power law but with

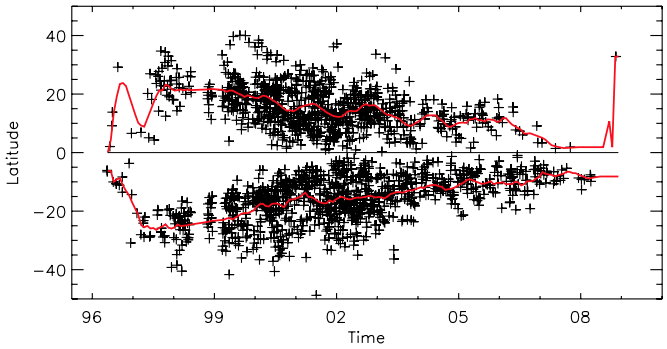

Figure 12. Distribution of AR location with time obtained from the automatic detection system. It appears similar to the classical Butterfly Diagram. The two red lines show the flux-weighted geometric center of ARs per Carrington rotation for the two solar hemispheres.

(A color version of this figure is available in the online journal.)

significant deviation at the two extreme ends. In the high end, which is of the largest ARs, the distribution decreases much faster than the power law, which could be fit by a softer power law, or exponential, or even Gaussian. In the lower end, the precipitous drop is likely caused by the detection cutoff of the auto system. The vast majority of detected features, starting from 2.8×10^{20} Mx (point A in the figure) and ending at 6.3×10^{22} Mx (point B), can be well fitted by a power-law function, described as

$$\frac{dN(F)}{d\log F} = 1.46 \times 10^3 \times F^{-0.630}, \quad (7)$$

where F again denotes the unsigned AR magnetic flux in units of 10^{20} Mx. The bin size of the distribution with respect to $\log F$ is again 0.05. In this case, the unreduced χ^2 test of the goodness of fit is only 0.146, a value that strongly supports the null hypothesis of the power-law function.

This result is consistent with that of recent work by Parnell et al. (2009), who found a power-law distribution of solar magnetic features over more than five decades in flux, expanding from large ARs to small bipolar regions in the quiet Sun. The only exception is that our data show a strong deviation from the dominant power law for those largest ARs, i.e., with flux larger than 6.3×10^{22} Mx. The change of AR distributions, from log-normal to power law when the controlling parameters are lowered, suggests important effects of computational bias on analysis results. One has to consider the bias when the results are discussed in scientific context.

4.4. Active Region Bands: Butterfly Diagram and Drifting Velocity

In this section, we investigate the distribution of AR locations on the Sun and the variation with the solar cycle. Figure 12 shows the heliographic latitude (Y-axis) of the 1730 ARs obtained by the automated method with respect to their time crossing the central meridian. The figure reproduces the classical Butterfly Diagram of solar ARs (Spörer's Law), which is caused by the emerging locations of ARs progressively drifting toward the equator, the sign of deep meridional flow of solar dynamo (Hathaway et al. 2003). The equatorward drifting motion from the two hemispheres is approximately symmetric to the equator, but not exact (more discussion later). As shown in the figure, almost all ARs are below 40° latitude, with a few exceptions in the southern hemisphere. The AR maximum latitudes in the northern and southern hemispheres are 40° and 48°, respectively.

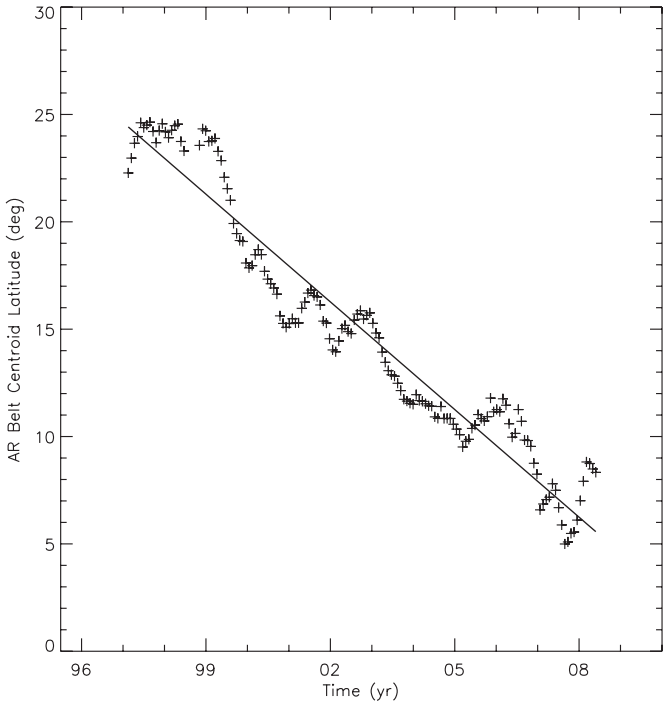

Figure 13. Latitudinal migration of AR bands, the so-called Spörer's Law. Each plus symbol represents the latitude of the geometric centroid of all ARs on each Carrington rotation; latitudes are folded. The black line is a linear fit to the data points, which measures the average drifting speed of the bands over the solar cycle.

ARs from the 23rd solar cycle started to emerge at high latitudes (∼30°) in late 1996. As the cycle progressed, not only did the emerging rate of ARs (as indicated by the number of ARs per Carrington rotation) increased, but also the latitudinal zones or bands of emerging ARs grew wider. Further, the centroids of the AR bands (indicated by the red lines) were drifting continuously toward the equator. AR bands were as wide as 30° during the rising phase of the solar cycle. In the middle of 2002, AR bands started to narrow down, and the number of emerging ARs also started to decrease. On the other hand, the centroids of the bands continued to move toward the equator. ARs completely disappeared in both hemispheres in the middle of 2008, and the absence of ARs was even earlier (by more than half year) in the northern hemisphere.

To quantify the drifting motion of the AR bands, we plot and analyze the centroid locations with respect to the solar cycle in Figure 13. The centroid location is derived from the mean latitude of all ARs per Carrington rotation weighted by magnetic flux. In order to have a clean representation of the drifting, we made an effort to use only ARs that emerged in Cycle 23. In the beginning of the cycle, we have removed the low-latitude ARs from Cycle 22, and in the end of the cycle, removed the high-latitude ARs from Cycle 24. Further, we consider only the folded latitudes from the equator, without differentiating between the northern and southern hemispheres. To the first order of approximation, the overall drifting can be described by the straight line shown in the figure. The line is the linear fit to the data points. The linear fitting formula is

$$\delta = 24.^{\circ}4 - 0.^{\circ}137 N_{\rm CR}, \quad (8)$$

where δ is the latitude in units of degree and $N_{\rm CR}$ is the number of Carrington rotation elapsed from the first data point in the line. The first cycle-23 AR appeared in CR 1919 and the last

cycle-23 AR in CR 2070. The fitting line shows that the AR band centroid started at about 25° and ended at about 5°. The average drifting rate is 0.°137/CR, with a standard deviation of 0.°003/CR. Converting to the drifting speed measured in years, since one Carrington rotation is 27.2753 days, and one year is 365.25 days, we obtain

$$V_d = 1.^{\circ}83 \pm 0.^{\circ}04\,\mathrm{yr}^{-1}. \quad (9)$$

In linear terms, the drifting speed is

$$V_d = 0.708 \pm 0.015\,\mathrm{m\ s}^{-1}. \quad (10)$$

We emphasize that this drifting speed or rate is an average value over the whole solar cycle. As seen from Figure 13, there are certain deviations from the oversimplified straight line. There seems a plateau for two years from 1997 to 1999. This plateau is followed by a speedy drifting much faster than the average speed. In at least four occasions, there appears a short period of "reverse" drifting, i.e., the centroid is moving away from the equator rather than toward it. The short period of "reverse" drifting is probably caused by short bursts of ARs emerging in usually high latitudes.

Earlier studies suggested that the drifting motion be fitted by a quadratic motion, based on historic sunspot data (Li et al. 2001; Hathaway et al. 2003). The quadratic fitting indicates that the drifting motion has a large velocity in the beginning of the solar cycle, gradually slow down, and eventually goes to zero at the end of the cycle. However, our data do not show such a gradually slowing trend. Instead, it seems to be a linear drifting on average, superposed with intermittent bursts of reverse drifting.

Finally, we want to mention the observation of the north–south asymmetry of the AR distribution. While the Butterfly Diagram and the drifting motion appear largely symmetric, there is also a noticeable asymmetry. Among the 1730 ARs obtained from our automated method, 938 of them are from southern hemisphere, while 792 are from the northern hemisphere. Therefore, over the whole solar cycle, there are about 18% more ARs appearing in the southern hemisphere than in the northern hemisphere. This south-over-north asymmetry mostly occurred during the declining phase of the solar cycle. In terms of magnetic flux, there is in total 1.55×10^{25} Mx that emerged in the southern hemisphere, while 1.33×10^{25} Mx emerged in the northern hemisphere; these numbers correspond to an asymmetry of about 17%. Similar north–south asymmetry has been reported earlier (Temmer et al. 2002; Zharkov & Zharkova 2006). The cause of this asymmetry is not understood. An investigation of the total budget of magnetic flux in the hemispheres requires knowledge of smaller magnetic regions, such as ephemeral regions. We will investigate the asymmetry issue in more detail in a separate work.

5. SUMMARY

We have developed a computational software system to automate the process of identifying and characterizing solar ARs. Based on the definition that an AR is an extended area of relatively strong magnetic fields, the system is built upon the image processing methods of morphological analysis and intensity thresholding. It consists of four functional modules: (1) intensity segmentation to obtain kernel pixels, (2) the morphological opening operation to erase small kernels, which effectively remove ephemeral regions and magnetic fragments in decayed ARs, (3) region growing to extend kernels to full AR size, and (4) the morphological closing operation to merge/group regions with a small spatial gap. Corresponding to the four modules, there are four controlling parameters (I_K, S_O, I_A, S_C) which effectively determine the detection result associated with a particular computational bias.

The system is tested and applied to high-resolution Carrington synoptic magnetograms constructed from MDI images from 1996 to 2008, and validated against the NOAA AR catalog. To reasonably reproduce the NOAA AR catalog, the controlling parameters are found to be (250 G, 10 Mm, 50 G, and 10 Mm). The average true positive rate, one metric to measure the relative success of the system, is 73.8%. The average false positive rate, the metric to measure the false-alarming error of the system, is found to be at 15.3%. Further, we find that the rate of compound ARs, each of which contains multiple NOAA ARs, is 13.0%.

When the detection thresholds are set higher, the number of detected ARs decreases, and both the true positive rate and the false positive rate also decrease. On the other hand, when the detection thresholds are chosen low, the number of detected regions increases, and both the true positive rate and the false positive rate increase. Therefore, each instance of detection has its own computational bias. However, the detection remains objective. Human operators create subjective errors, which tend to be random and cannot be repeated. It makes the situation worse when multiple inspectors are involved to create a long-term data catalog, such as the daily sunspot number (Hathaway 2010). On the other hand, the computational bias is a controlled one, which can be repeated at a later time and/or by other users. Further, our study demonstrates that the automated system can be tuned, based on the scientific needs, to different types of detections, i.e., large ARs only or regions including smaller magnetic features.

With the "optimized" controlling parameters, the automated system finds 1730 ARs from 1996 and 2008. We further calculate the basic geometric and flux parameters of these regions, based on which we revisit several well-known solar AR properties related to solar cycle and solar dynamo. Our AR number, counted per Carrington rotation, varies closely in phase with the NOAA AR number over the whole solar cycle. However, we find that, while the NOAA AR number shows double peaks of similar strength during the solar maximum period, one in early 2000, and the other in late 2001, the magnetic flux (and geometric area as well) shows one outstanding peak in late 2001. Mean and maximum magnetic flux of individual ARs are 1.67×10^{22} Mx and 1.97×10^{23} Mx, while that per Carrington rotation are 1.83×10^{23} Mx and 6.96×10^{23} Mx, respectively. We further find that the geometric size of an AR measured from magnetograms is about 15 times on average as large as that of the corresponding sunspot measured in white light images.

In terms of AR frequency distribution, we find that both area size and magnetic flux distributions could fit into a log-normal function, albeit the large χ^2 value. The log-normal function is consistent with the result by Bogdan et al. (1988), who argued that ARs were caused by the fragmentation process through the convection zone, thus favoring the global dynamo model. However, when we decrease the detection thresholds to (150 G, 6 Mm, 50 G, and 6 Mm), the frequency distribution of detected regions turns out to be a power-law function for most of the data range (from 2.8×10^{20} Mx to 6.3×10^{22} Mx) except for the largest ARs. Parnell et al. (2009) recently found a power-law distribution of solar magnetic features over more

than five decades in flux, including ARs. We argue that the distribution of solar magnetic features from large to small may have two components: (1) a log-normal component for large ARs and (2) a power-law component for small ARs and other smaller features including ephemeral regions, magnetic fragments in decayed ARs, and magnetic elements in the quiet-Sun network. The magnetic features of the second component are known to be subject to near-surface dynamo (Schrijver et al. 1997). However, the theoretical fragmentation-merging model of Schrijver et al. (1997) produces an exponential function of distribution. Nevertheless, the two-component distribution discussed above implies that both global and near-surface dynamo models operate in the Sun to produce the observed magnetic features.

We have also investigated Spörer's Law or the equatorward drifting motion of the AR bands. We find that the motion can be described by a linear function superposed by intermittent reverse driftings. The constant drifting speed is $1.^{\circ}83 \pm 0.^{\circ}04\ \mathrm{yr}^{-1}$ or $0.708 \pm 0.015\ \mathrm{m\ s}^{-1}$. Note that the drifting motion was described before by a quadratic model (Li et al. 2001; Hathaway et al. 2003). Since different data are used, we do not discuss further the validity of these models. Our data also prove the north–south hemispheric asymmetry (in terms of AR number, area, and magnetic flux) that has been reported earlier (Temmer et al. 2002; Zharkov & Zharkova 2006).

In short, we conclude that the automated detection system is not only efficient and objective, but also effective in addressing scientific issues. The computational biases associated with the system can be understood and repeated, which can be further used to tune the system to fit different detection purposes. In particular, it would be interesting to detect and study small magnetic features. Even though the system is applied to Carrington synoptic magnetograms, it can be easily adopted to study snapshot magnetograms. When a tracking module is implemented, one can study the evolutions of individual ARs, and further predict the possibility of these ARs producing flares and CMEs.

We thank the anonymous referee for valuable comments. We acknowledge the usage of *SOHO*/MDI data. *SOHO* is a project of international cooperation between ESA and NASA. J.Z. is supported by NASA grants NNG07AO72G and NSF ATM-0748003. Y.M.W. is supported by 973-key-project 2006CB806304 and NSFC 40525014 of China.

REFERENCES

Abramenko, V. I., Yurchyshyn, V. B., Wang, H., Spirock, T. J., & Goode, P. R. 2003, ApJ, 597, 1135
Aschwanden, M. J. 2010, Sol. Phys., 262, 235
Barnes, G., & Leka, K. D. 2008, ApJ, 688, L107
Berger, T. E., & Lites, B. W. 2003, Sol. Phys., 213, 213
Bogdan, T. J., Gilman, P. A., Lerche, I., & Howard, R. 1988, ApJ, 327, 451
Boursier, Y., Lamy, P., Llebaria, A., Goudail, F., & Robelus, S. 2009, Sol. Phys., 257, 125
Colak, T., & Qahwaji, R. 2008, Sol. Phys., 248, 277
Colak, T., & Qahwaji, R. 2009, Space Weather, 7, 6001
Conlon, P. A., Gallagher, P. T., McAteer, R. T. J., Ireland, J., Young, C. A., Kestener, P., Hewett, R. J., & Maguire, K. 2008, Sol. Phys., 248, 297
Falconer, D. A., Moore, R. L., & Gary, G. A. 2002, ApJ, 569, 1016
Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 661, L109
Hagenaar, H. J., Schrijver, C. J., & Title, A. M. 2003, ApJ, 584, 1107
Hale, G. E., Ellerman, F., Nicholson, S. B., & Joy, A. H. 1919, ApJ, 49, 153
Harvey, K. L., & Zwaan, C. 1993, Sol. Phys., 148, 85
Hathaway, D. H. 2010, Living Rev. Sol. Phys., 7, 1
Hathaway, D. H., Nandy, D., Wilson, R. M., & Reichmann, E. J. 2003, ApJ, 589, 665
Hathaway, D. H., Wilson, R. M., & Reichmann, E. J. 1999, J. Geophys. Res., 104, 22375
Howard, R. F. 1989, Sol. Phys., 123, 271
Ireland, J., Young, C. A., McAteer, R. T. J., Whelan, C., Hewett, R. J., & Gallagher, P. T. 2008, Sol. Phys., 252, 121
Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. 2006, ApJ, 644, 1273
Lawrence, J. K., Ruzmaikin, A. A., & Cadavid, A. C. 1993, ApJ, 417, 805
Leka, K. D., & Barnes, G. 2003, ApJ, 595, 1277
Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173
Li, K. J., Yun, H. S., & Gu, X. M. 2001, AJ, 122, 2115
Maunder, E. W. 1904, MNRAS, 64, 747
McAteer, R. T. J., Gallagher, P. T., & Ireland, J. 2005, ApJ, 631, 628
McIntosh, P. S. 1990, Sol. Phys., 125, 251
McKinnon, J. A., & Waldmeier, M. (ed.) 1987, Sunspot Numbers, 1610–1985: Based on The Sunspot Activity in the years 1610–1960 (Boulder, CO: World Data Center A for Solar-Terrestrial Physics)
Meunier, N. 1999, ApJ, 515, 801
Olmedo, O., Zhang, J., Wechsler, H., Poland, A., & Borne, K. 2008, Sol. Phys., 248, 485
Parnell, C. E., DeForest, C. E., Hagenaar, H. J., Johnston, B. A., Lamb, D. A., & Welsch, B. T. 2009, ApJ, 698, 75
Qu, M., Shih, F. Y., Jing, J., & Wang, H. 2004, Sol. Phys., 222, 137
Qu, M., Shih, F. Y., Jing, J., & Wang, H. 2005, Sol. Phys., 228, 119
Robbrecht, E., & Berghmans, D. 2004, A&A, 425, 1097
Scherrer, P. H., et al. 1995, Sol. Phys., 162, 129
Schrijver, C. J. 2007, ApJ, 655, L117
Schrijver, C. J., Title, A. M., van Ballegooijen, A. A., Hagenaar, H. J., & Shine, R. A. 1997, ApJ, 487, 424
Tang, F., Howard, R., & Adkins, J. M. 1984, Sol. Phys., 91, 75
Temmer, M., Veronig, A., & Hanslmeier, A. 2002, A&A, 390, 707
Tran, T., Bertello, L., Ulrich, R. K., & Evans, S. 2005, ApJS, 156, 295
Turmon, M., Pap, J. M., & Mukhtar, S. 2002, ApJ, 568, 396
Ulrich, R. K., Bertello, L., Boyden, J. E., & Webster, L. 2009, Sol. Phys., 255, 53
Wang, Y., & Zhang, J. 2008, ApJ, 680, 1516
Wang, Y., et al. 2010, ApJ, 717, 973
Wolf, R. 1861, MNRAS, 21, 77
Zharkov, S. I., & Zharkova, V. V. 2006, Adv. Space Res., 38, 868
Zharkov, S. I., Zharkova, V. V., & Ipson, S. S. 2005, Sol. Phys., 228, 377
Zharkova, V. V., Aboudarham, J., Zharkov, S., Ipson, S. S., Benkhalil, A. K., & Fuller, N. 2005, Sol. Phys., 228, 361