Solar Phys (2008) 248: 323–337
DOI 10.1007/s11207-007-9069-4

# A Comparison of Feature Classification Methods for Modeling Solar Irradiance Variation

**H.P. Jones · G.A. Chapman · K.L. Harvey · J.M. Pap · D.G. Preminger · M.J. Turmon · S.R. Walton**

Received: 4 May 2007 / Accepted: 4 October 2007 / Published online: 10 November 2007


**Abstract** Physical understanding of total and spectral solar irradiance variation depends upon establishing a connection between the temporal variability of spatially resolved solar structures and spacecraft observations of irradiance. One difficulty in comparing models derived from different data sets is that the many ways for identifying solar features such as faculae, sunspots, quiet Sun, and various types of "network" are not necessarily consistent. To learn more about classification differences and how they affect irradiance models, feature "masks" are compared as derived from five current methods: multidimensional histogram analysis of NASA/National Solar Observatory/Kitt Peak spectromagnetograph data, statistical pattern recognition applied to SOHO/Michelson Doppler Imager photograms and magnetograms, threshold masks allowing for influence of spatial surroundings applied to NSO magnetograms, and "one-trigger" and "three-trigger" algorithms applied to California State University at Northridge Cartesian Full Disk Telescope intensity observations. In general all of the methods point to the same areas of the Sun for labeling sunspots and active-region faculae, and available time series of area measurements from the methods correlate well with each other and with solar irradiance. However, some methods include larger label sets, and there are important differences in detail, with measurements of sunspot area

K.L. Harvey and S.R. Walton are deseased, to whom this paper is dedicated.

H.P. Jones (✉)
National Solar Observatory, P.O. Box 26732, Tucson, AZ 85726, USA
e-mail: hjones@nso.edu

G.A. Chapman · D.G. Preminger · S.R. Walton
Department of Physics and Astronomy, California State University Northridge, Northridge, CA, USA

K.L. Harvey
Solar Physics Research Corporation, Tucson, AZ, USA

J.M. Pap
Goddard Earth Science and Technology Institute/UMBC, NASA's Goddard Space Flight Center, Greenbelt, MD, USA

M.J. Turmon
NASA's Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

differing by as much as a factor of two. The methods differ substantially regarding inclusion of fine spatial scale in the feature definitions. The implications of these differences for modeling solar irradiance variation are discussed.

**Keywords** Solar irradiance · Sunspots · Active regions · Photosphere · Data analysis

## 1. Introduction

Solar irradiance variation has been measured with high relative accuracy with spacecraft radiometers since late 1978 (Pap, 2003; Fröhlich, 2006; Lean, 2001). These instruments have shown that the Sun's total irradiance decreases in the presence of sunspots and increases in the presence of continuum faculae (Willson *et al.*, 1981). The most important discovery of these measurements was that total irradiance changes over the solar cycle with an amplitude of 0.1%, being higher during maximum activity conditions (Willson and Hudson, 1988). However, despite convincing demonstrations (Foukal and Lean, 1988; Chapman, Cookson, and Dobias, 1997; Fligge, Solanki, and Unruh, 2000; Preminger, Walton, and Chapman, 2002; Krivova *et al.*, 2003; Wenzler *et al.*, 2006) that sunspots and faculae are the dominant sources of total solar irradiance variation on time scales of up to several years, many questions remain.

One major problem (Pap *et al.*, 2002) is that the uncertainty in the absolute calibration of the various radiometers exceeds the observed magnitude of solar variability. Since there are gaps in the observing periods of various spacecraft, there is no unambiguous way to ensure that attempts to combine the various observations into a long-term continuous record do not confuse errors in absolute calibration and unmeasured instrumental degradation with true solar variability. Indeed, major reconstructions of spacecraft radiometer measurements (Fröhlich, 2006; Fröhlich and Lean, 1998; Willson and Mordvinov, 2003; Lean *et al.*, 2005) differ in important details, such as whether the average irradiance at solar minimum is approximately constant from cycle to cycle. Moreover, although the reconstructions agree that total irradiance at the maxima of cycles 21, 22, and 23 is about the same, many observations (de Toma *et al.* 2001, 2004; Pap *et al.* 2002) suggest that photospheric magnetic flux, the contribution of photospheric faculae, and sunspot number and area are less at the maximum for cycle 23 than they were at the corresponding phase of cycle 22. Finally, the explained variance ($\approx 80\%$) of even the best long-term models (Wenzler *et al.*, 2006), while consistent with ground-based measurement errors, easily allows for global or other sources of irradiance variation outside the sunspot–faculae model. The uncertainties are not large over the period spanned by spacecraft observations, but they can lead to very different interpretations of the historical sunspot record (Fligge and Solanki, 2000; Lean, Beer, and Bradley, 1995; Hoyt and Schatten, 1993). These differences in turn lead to large uncertainties in the solar forcing of climate variations and complicate interpretation of the magnitude of anthropogenic causes of global terrestrial warming.

One essential ingredient for improving our understanding of solar irradiance variation in the era of spacecraft observations is the accurate identification of narrowband or broadband (Foukal *et al.*, 2004) bright and dark solar features and their relation to historically observed indicators of solar activity. Although solar physicists can easily identify many kinds of solar features, visual agreement on precise spatial structure is data dependent and generally difficult or impossible to attain, especially for intrinsically low-contrast features such as faculae. Such difficulties together with the large volume of available data have motivated many automatic or computer-aided feature recognition methods. In the following sections, we examine

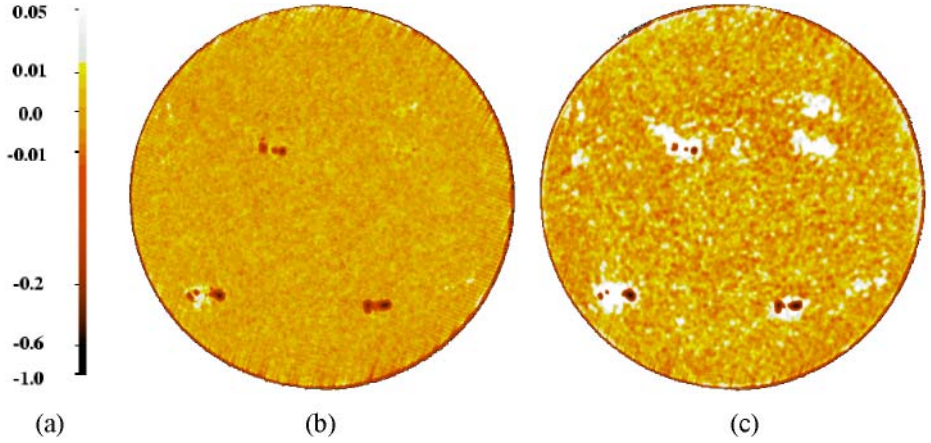


**Figure 1** (b) Red and (c) Ca K-line images from San Fernando Observatory on 9 September 1997. Mapping of intensity contrast into color is shown by the color bar (a) and applies to all intensity contrast images.

and compare several published methods for identifying solar features. We give no absolute answers on the quality or correctness of any of the methods but instead develop methodology for quantitative comparison and discuss several important issues pertaining to feature identification.

## 2. Observations and Feature Identification Methods

The feature identification methods we discuss here label areas of interest on solar images obtained with different instruments. The observations from all these instruments are preprocessed to remove artifacts and solar characteristics such as center-to-limb variation as discussed in the cited references. We will not duplicate these discussions but note that this preprocessing is crucial for all the methods since residual systematics and artifacts are a source of error that will produce differences between labelings even if the methods would otherwise agree perfectly.

The "one-trigger"(1T) method (Chapman *et al.*, 1992) uses simple thresholds to distinguish dark sunspots (ss) in red continuum contrast images ($i_c$) and bright faculae (f) in Ca K core-wing images ($i_{cw}$) from the "quiet-Sun" background; observations are obtained with the Cartesian Full-Disk Telescope-1 (CFDT1) at the San Fernando Observatory (SFO) of California State University Northridge (CSUN). As in all contrast images discussed in this paper, contrast is determined by comparing the actual observations to the mean center-to-limb variation. CSUN/SFO images from 9 September 1997 used for comparative analysis in this paper are shown in Figure 1. Sunspots and faculae derived by this method have been compared extensively with spacecraft observations of irradiance variation in a number of studies (*e.g.*, Chapman, Cookson, and Dobias, 1996, 1997). The "three-trigger" (3T) algorithm (Preminger, Walton, and Chapman, 2001) is a more sophisticated application of contrast thresholds that requires that the contrast of three contiguous pixels exceed a prespecified value to initiate a search for other neighboring pixels that deviate by more than a (possibly less restrictive) second threshold. The authors show evidence that the method identifies real larger feature areas than the "one-trigger" method without much confusion with instrumental noise (Preminger, Walton, and Chapman, 2002).

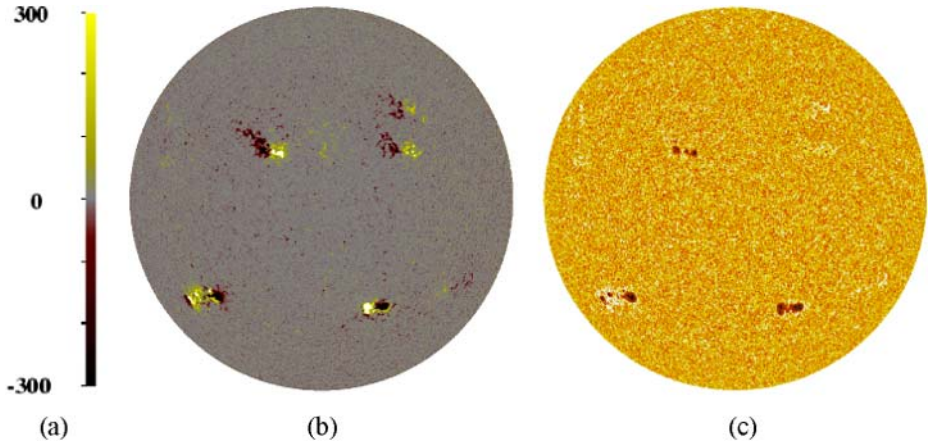


**Figure 2** (b) SOHO/MDI magnetogram and (c) intensity contrast image on 9 September 1997. Mapping of line-of-sight magnetic flux to color is shown by the color bar (a) and applies to both MDI and SPM magnetograms. Intensity color mapping is as in Figure 1a.

Turmon, Pap, and Mukhtar (2002) avoid the use of thresholds by applying a statistical method to SOHO Michelson Doppler Imager (MDI) line-of-sight magnetograms ($B_{\parallel}$) and continuum intensity ($i_c$) contrast images (photograms) taken close together in time. Models for "class-conditional" probabilities that a pixel will have observed characteristics given that it is a member of a specified feature class are developed from distributions of magnetic flux and intensity in an independently segmented training set of images. These class-conditional models are inverted with Bayes' rule to produce probability images for sunspots, faculae, and quiet Sun given the actual observed properties of images other than the training set. The MDI images analyzed here are shown in Figure 2. A simulated-annealing algorithm then searches for the image labeling with the maximum global (over the entire image) *a posteriori* probability. Contiguity of pixels is modeled in the Bayesian prior probability. We refer to this technique as the maximum likelihood (ML) method.

Harvey and White (1999) apply thresholds sequentially to strictly cotemporal and cospatial line-of-sight magnetograms and continuum contrast images (868.8 nm). The observations were obtained with the NASA/National Solar Observatory (NSO) Spectromagnetograph (SPM) at the NSO/Kitt Peak Vacuum Telescope (KPVT), and they are also the basis for the Jones *et al.* (2000, 2003) method (Figure 3). The Harvey and White algorithm begins with separation of "magnetized" and "unmagnetized" regions with a threshold of magnetic flux comparable to the noise level of the SPM. In the magnetized regions, they first extract sunspot umbrae and penumbrae using thresholds in both magnetic flux and intensity. They then proceed to use polarity inversion, fill factors, and thresholds to extract active regions, decaying regions, and various kinds of network. They account for contiguity of pixels in each feature class and smooth with Gaussian filters at various stages of the process. We use a summary classification of their method that labels sunspots (ss; umbrae + penumbrae), active regions (ar), decaying regions (dr), enhanced network (en), and quiet network (qn; weak + quiet network in the nomenclature of Harvey and White, 1999). We refer to this technique as the sequential thresholds (ST) method.

Jones *et al.* (2000, 2003) apply thresholds to isolate potentially interesting subdomains of multidimensional histograms of SPM observations [line-of-sight flux, intensity contrast, equivalent width contrast ($w$), line depth contrast ($d$), unipolarity ($u$), and heliocentric an-

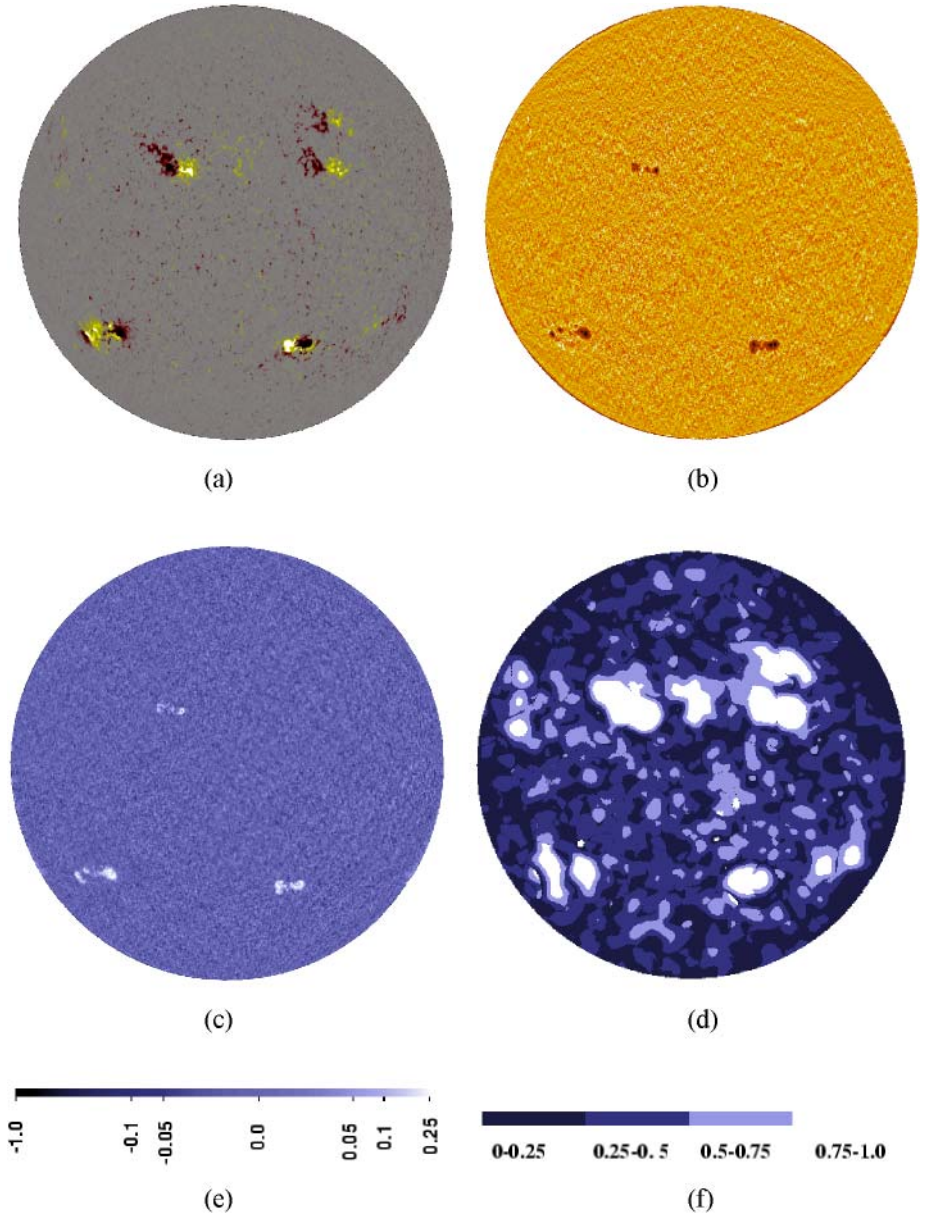


**Figure 3** NASA/NSO Spectromagnetograph images on 9 September 1997: (a) magnetogram, (b) intensity contrast, (c) equivalent width contrast, and (d) unipolarity. Magnetogram color mapping is as in Figure 2a; intensity color mapping is as in Figure 1a. Equivalent width contrast mapping is shown by color bar (e), and unipolarity mapping by color bar (f).

gle]. Unipolarity for each pixel is calculated as the absolute value of the fractional difference between positive and negative flux over approximately a 30 × 30 arcsecond surrounding area. The authors apply factor analysis over extensive time series of these observations to

**Table 1** Characteristics of the one-trigger (1T), three-trigger (3T), maximum likelihood (ML), sequential threshold (ST), and histogram factor (HF) feature identification methods. Abbreviations for observed quantities and feature definitions are described in the text.

| | Method | | | | |
|---|---|---|---|---|---|
| | 1T | 3T | ML | ST | HF |
| Instrument | SFO/CFDT1 | | SOHO/MDI | NASA/NSO SPM | |
| $\lambda$ | $672.3 \pm 10$ nm<br>$393.4 \pm 1$ nm | | 676.8 nm (Ni I) | 868.8 nm (Fe I) | |
| Observables | $i_c$(672.3), $i_{cw}$(Ca II K) | | $B_\parallel$, $i_c$(676.8) | $B_\parallel$, $i_c$(868.8) | $B_\parallel$, $i_c$, $w$, $d$, $u$ |
| Features | ss (672.3 nm)<br>f (Ca II K)<br>qs | | ss<br>f<br>qs | ss<br>ar<br>dr<br>en<br>qn | ss<br>uni<br>mx<br>wk<br>unas |
| Comparison with total solar irradiance | | | | | |
| Interval | 1988 – 2005 | 1988 – 2005 | 1996 – 2000 | NA | 1992 – 2000 |
| $R^2$ | 0.79 | 0.79 | 0.85 | NA | 0.77 |

find the orthogonal (uncorrelated) linear combinations of the original subdomains that explain most of the variance of the SPM observations. Three factors (sunspots – umbrae + penumbrae; strong unipolar + bright limb regions, and magnetically strong mixed polarity regions) account for most of the SPM variance but only the first two are strongly correlated with total irradiance variation. Although there is no strictly accurate way to establish a one-to-one correspondence with factor values and pixel position, the gross factor structure is clear enough that a reasonably accurate spatial labeling can be established. Here, we label pixels as weak-field (wk), (strong) mixed polarity (mx), unipolar + bright limb (uni), or sunspots (ss) and leave an unassigned category (unas) because the histogram subdomains defined by these labelings do not span the entire variable space. We refer to this technique as the histogram factor (HF) method.

The observations from SFO and NSO were rotated differentially to correspond to the time of the MDI magnetogram, and common features in the intensity images of all three instruments as well as on the NSO and MDI magnetograms were visually identified and their positions were marked. The SFO and NSO images were then "warped" by using the IRAF geomap and geotran procedures to complete detailed spatial registration with the MDI observations. Corresponding transformations were applied to the feature labelings, using nearest neighbor interpolation. Errors associated with spatial registration will appear in the comparisons that follow, but we believe they are negligible compared to the differences arising from the different methods of feature classification.

For convenience in the following discussion, salient features of the five methods are summarized in Table 1. Also included is multiple $R^2$ for linear regressions of TSI on feature statistics over time. Harvey and White (1999) made no such comparison for ST. Despite

their many differences, all the methods correlate comparably with TSI and perform better for shorter periods.

## 3. Comparison of Feature Labelings

The image labelings from the various methods are shown in Figure 4; many obvious differences are apparent. For example, the labeling schemes for the ST and the HF methods are more complicated than the simple (quiet Sun, faculae, and sunspot) classification for the 1T, 3T, and ML methods. The HF labelings, which allow for contiguity only through the use of the unipolarity information, show many more fine-scale structures than the other methods.

More quantitative cross-tabulations of the ten possible pairs of label comparisons are shown as "confusion matrices" in Tables 2–5. Each table shows, from a pixel-by-pixel comparison of two segmentations, the fraction of pixels labeled by both the indicated row and column features. For example, the first column and second row of Table 2 indicates that 0.021676 of the total number of pixels were labeled quiet Sun by the 1T method but faculae by the 3T method. Diagonal entries show the fraction of pixels that were classified the same by both methods, and perfect agreement, where the categories are putatively the same, would be shown with nonzero values only along the diagonals. As a figure of merit, we also show (along with asymptotic standard errors, since large numbers of pixels are involved even for sunspots) Cohen's $\kappa$ (Cohen, 1960), which gives an assessment of agreement based on the assumption that expected random values in the tables are the simple products of the corresponding marginal probabilities (the row and column totals in Tables 2–5). If $O_{i,j}$ are the observed entries and $E_{i,j} \equiv (\sum_j O_{i,j}) \times (\sum_i O_{i,j})$ are the expected values,

$$\kappa = \left(\sum_i O_{i,i} - \sum_i E_{i,i}\right) \Big/ \left(1 - \sum_i E_{i,i}\right). \tag{1}$$

By comparing with expected random values, $\kappa$ avoids giving an exaggerated impression of agreement where, as is the case here, one class (quiet Sun) dominates the others. For perfect agreement, $\sum_i O_{i,i} = 1$, so that $\kappa = 1$, whereas $\kappa = 0$ for random agreement. Given the observed marginals, however, the maximum possible value ($\kappa_{\max}$) is achieved when the minimum of the two marginals for category $i$ is substituted for $O_{i,i}$ in Equation (1). In Tables 2–5 we also show $\kappa/\kappa_{\max}$, which gives a less conservative measure of agreement, particularly when the labelings of one method are largely inclusive of the labelings of the other. A requirement is that the categories be the same between two methods. We discuss, in the following, groupings of the categories for the HF and ST methods that allow approximate comparison with each other and with the (quiet Sun, faculae, and sunspot) classification of 1T, 3T, and ML. We use Cohen's $\kappa$ only as a convenient summary metric, which at least approximately orders the quality of the various comparisons, and suggest this measure may also be useful for such purposes as tuning parameters of a given segmentation method. More rigorous statistical inferences based on the metric are beyond the scope of this paper.

There are of course many ways in which differences in segmentations can arise, and there are other ways of presenting the information. Inspection of Figure 4 shows that for the most part the methods agree on the location of sunspots and faculae or the approximate equivalents in the ST and HF methods. The differences arise in the details of the segmentation definitions. We have also examined but, for clarity and compactness, do not show ten detailed pixel-by-pixel comparison images that show the spatial structure of the categories summarized by Tables 2–5. Although interesting details are apparent, those relevant to the

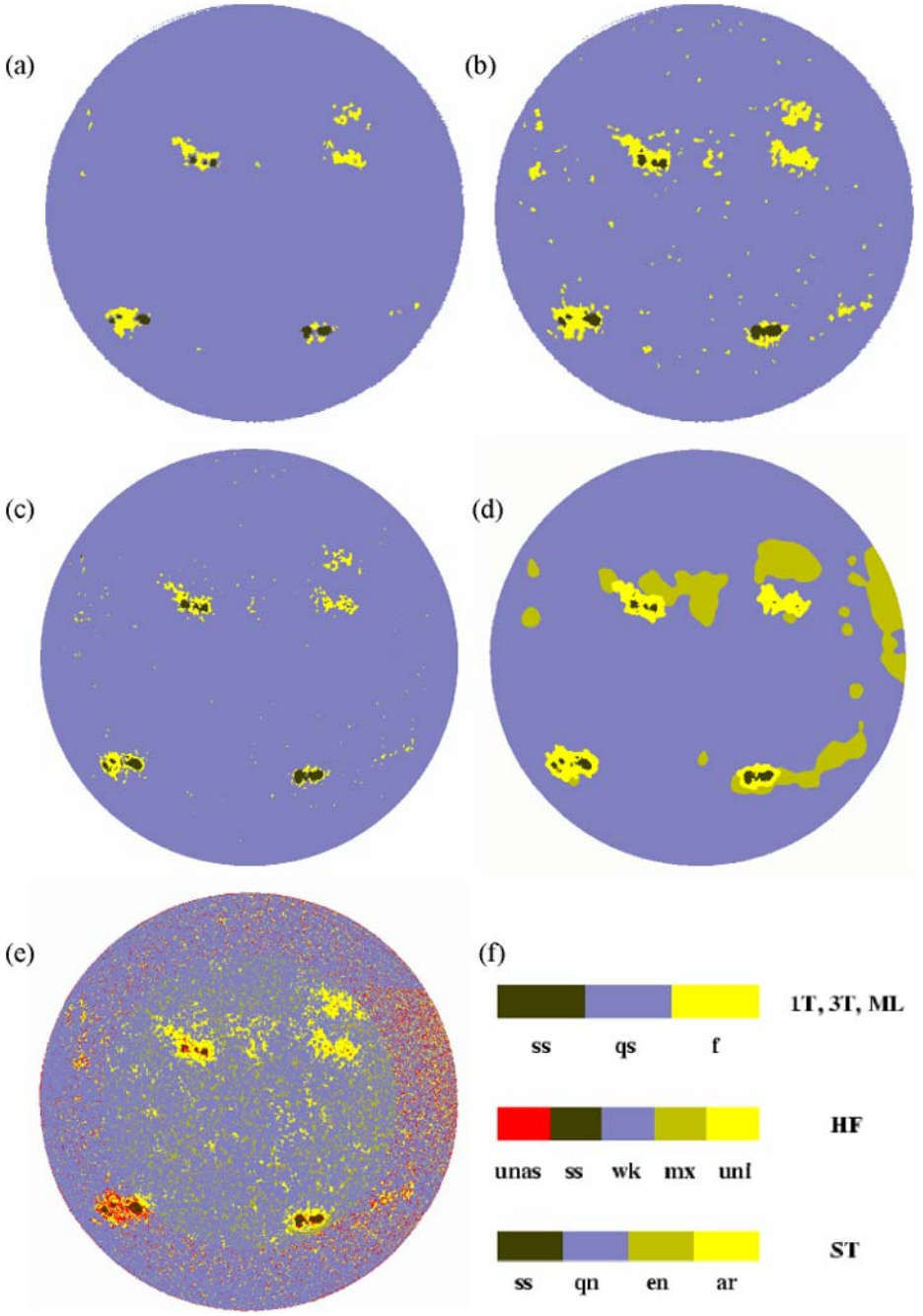


**Figure 4** Feature labelings by the (a) 1T, (b) 3T, (c) ML, (d) ST, and (e) HF methods. Color tables are shown in panel (f). Feature abbreviations are described in the text.

**Table 2** Fraction of one-trigger labelings for each label category of the three-trigger, maximum likelihood, histogram factor, and sequential thresholds methods. Also shown for each comparison are Cohen's $\kappa$ (see text) $\pm$ the asymptotic standard error and $\kappa/\kappa_{\max}$.

| | One-trigger | | | |
|---|---|---|---|---|
| | Quiet Sun | Faculae | Sunspot | Total |
| Three-trigger | | | | |
| Quiet Sun | 0.959813 | 0.000000 | 0.000000 | 0.959813 |
| Faculae | 0.021676 | 0.013105 | 0.000000 | 0.034782 |
| Sunspot | 0.001725 | 0.000496 | 0.003185 | 0.005405 |
| Total | 0.983214 | 0.013601 | 0.003185 | 1.000000 |
| $\kappa = 0.57 \pm 0.003;\ \kappa/\kappa_{\max} = 0.98$ | | | | |
| Maximum likelihood | | | | |
| Quiet Sun | 0.975289 | 0.005166 | 0.000022 | 0.980477 |
| Faculae | 0.006545 | 0.007922 | 0.000232 | 0.014699 |
| Sunspot | 0.001234 | 0.000631 | 0.002959 | 0.004825 |
| Total | 0.983068 | 0.013719 | 0.003212 | 1.000000 |
| $\kappa = 0.61 \pm 0.003;\ \kappa/\kappa_{\max} = 0.66$ | | | | |
| Histogram factor | | | | |
| Weak field | 0.813898 | 0.001625 | 0.000045 | 0.815568 |
| Strong mixed polarity | 0.052326 | 0.000066 | 0.000004 | 0.052396 |
| Unipolar + bright limb | 0.076339 | 0.009770 | 0.000243 | 0.086353 |
| Sunspot | 0.000479 | 0.000197 | 0.002178 | 0.002854 |
| Unassigned | 0.040053 | 0.002039 | 0.000737 | 0.042829 |
| Total | 0.983095 | 0.013697 | 0.003207 | 1.000000 |
| $\kappa = 0.21 \pm 0.002;\ \kappa/\kappa_{\max} = 0.70$ | | | | |
| Sequential thresholds | | | | |
| Quiet network | 0.884605 | 0.000286 | 0.000000 | 0.884891 |
| Enhanced network | 0.080937 | 0.002406 | 0.000000 | 0.083344 |
| Active region | 0.016362 | 0.010633 | 0.000412 | 0.027407 |
| Sunspot | 0.001086 | 0.000457 | 0.002815 | 0.004358 |
| Total | 0.982990 | 0.013782 | 0.003227 | 1.000000 |
| $\kappa = 0.56 \pm 0.003;\ \kappa/\kappa_{\max} = 0.81$ | | | | |

purposes of this paper can also be seen in Figure 4 and are reasonably summarized in the tables.

Table 2 compares the 1T labeling with each of the other methods. The 3T method, as discussed in more detail by Preminger, Walton, and Chapman (2001), finds more facular and sunspot pixels than the 1T method, with most of the additional area for both cases coming from areas identified as quiet Sun in the one-trigger labeling. Although the facular

**Table 3** Comparison of 3T with ML, HF, and ST as in Table 2.

| | Three-trigger | | | |
|---|---|---|---|---|
| | Quiet Sun | Faculae | Sunspot | Total |
| Maximum likelihood | | | | |
| Quiet Sun | 0.957087 | 0.023074 | 0.000316 | 0.980477 |
| Faculae | 0.002168 | 0.011478 | 0.001053 | 0.014699 |
| Sunspot | 0.000210 | 0.000531 | 0.004084 | 0.004825 |
| Total | 0.959464 | 0.035084 | 0.005452 | 1.000000 |
| $\kappa = 0.53 \pm 0.003;\ \kappa/\kappa_{\max} = 0.83$ | | | | |
| Histogram factor | | | | |
| Weak field | 0.805260 | 0.010199 | 0.000108 | 0.815568 |
| Strong mixed polarity | 0.051375 | 0.001016 | 0.000005 | 0.052396 |
| Unipolar + bright limb | 0.064918 | 0.020079 | 0.001355 | 0.086353 |
| Sunspot | 0.000055 | 0.000150 | 0.002649 | 0.002854 |
| Unassigned | 0.037920 | 0.003583 | 0.001326 | 0.042829 |
| Total | 0.959529 | 0.035028 | 0.005443 | 1.000000 |
| $\kappa = 0.32 \pm 0.002;\ \kappa/\kappa_{\max} = 0.56$ | | | | |
| Sequential thresholds | | | | |
| Quiet network | 0.879247 | 0.005644 | 0.000000 | 0.884891 |
| Enhanced network | 0.072215 | 0.011128 | 0.000000 | 0.083344 |
| Active region | 0.007634 | 0.018103 | 0.001670 | 0.027407 |
| Sunspot | 0.000181 | 0.000370 | 0.003807 | 0.004358 |
| Total | 0.959277 | 0.035245 | 0.005477 | 1.000000 |
| $\kappa = 0.62 \pm 0.002;\ \kappa/\kappa_{\max} = 0.71$ | | | | |

areas identified by the 1T and ML labelings are similar, only about half of the facular pixels identified by either method come from the same areas on the Sun; the remainder are labeled quiet Sun by the other method. The sunspot areas identified by the ML method are larger than those for the 1T method, with most of the difference coming from 1T quiet-Sun areas.

Comparison with both the HF and ST labelings is more complicated. A rough correspondence between the HF and 1T labelings can be made by combining the weak-field, strong mixed polarity, and unassigned categories as corresponding to quiet Sun and unipolar regions as corresponding to faculae. Indeed, in the Jones *et al.* papers, unipolar + bright limb regions correlated strongly and positively with total solar irradiance variation, as would be expected for faculae. As in the comparison with the ML method, however, there is then still considerable confusion between the quiet-Sun and facular regions. The HF method identifies a far larger area corresponding to unipolar regions than the facular area labeled by the 1T method, with most of the discrepancy arising in 1T quiet-Sun regions. Finally, the sunspot area labeled by the HF method is somewhat similar to but slightly smaller than the corresponding 1T area.

**Table 4** Comparison of ML with HF and ST labelings.

| | Maximum likelihood | | | |
|---|---|---|---|---|
| | Quiet Sun | Faculae | Sunspot | Total |
| Histogram factor | | | | |
| Weak field | 0.815227 | 0.000335 | 0.000086 | 0.815648 |
| Strong mixed polarity | 0.052096 | 0.000174 | 0.000007 | 0.052277 |
| Unipolar + bright limb | 0.073739 | 0.012047 | 0.000636 | 0.086421 |
| Sunspot | 0.000004 | 0.000147 | 0.002697 | 0.002848 |
| Unassigned | 0.039487 | 0.001939 | 0.001381 | 0.042807 |
| Total | 0.980552 | 0.014642 | 0.004806 | 1.000000 |
| $\kappa = 0.26 \pm 0.002;\ \kappa/\kappa_{\max} = 0.80$ | | | | |
| Sequential thresholds | | | | |
| Quiet network | 0.883910 | 0.000984 | 0.000038 | 0.884932 |
| Enhanced network | 0.080293 | 0.003061 | 0.000012 | 0.083365 |
| Active region | 0.016146 | 0.010440 | 0.000768 | 0.027353 |
| Sunspot | 0.000076 | 0.000254 | 0.004020 | 0.004350 |
| Total | 0.980425 | 0.014738 | 0.004837 | 1.000000 |
| $\kappa = 0.58 \pm 0.003;\ \kappa/\kappa_{\max} = 0.77$ | | | | |

**Table 5** Comparison of ST and HF labelings.

| | ST | | | | |
|---|---|---|---|---|---|
| | Q. Network | E. Network | Active region | Sunspot | Total |
| HF | | | | | |
| Weak field | 0.758463 | 0.053000 | 0.005294 | 0.000057 | 0.816813 |
| Mixed polarity | 0.049992 | 0.002112 | 0.000513 | 0.000000 | 0.052618 |
| Unipolar | 0.047937 | 0.020560 | 0.017854 | 0.000314 | 0.086664 |
| Sunspot | 0.000000 | 0.000000 | 0.000187 | 0.002679 | 0.002866 |
| Unassigned | 0.028542 | 0.007693 | 0.003505 | 0.001300 | 0.041039 |
| Total | 0.884933 | 0.083364 | 0.027353 | 0.004350 | 1.000000 |
| $\kappa = 0.31 \pm 0.002;\ \kappa/\kappa_{\max} = 0.64$ | | | | | |

Similarly, a rough correspondence between the ST and 1T labelings is seen if the ST quiet and enhanced network categories are combined and considered as quiet Sun. The ST active region area is roughly twice as large as the 1T facular area, with most of the difference coming from one-trigger quiet Sun, and the ST sunspot area is about 50% larger than that of the one-trigger case.

Table 3 shows similar comparisons for the 3T with the ML, HF, and ST methods. Here, the 3T facular area is more than twice that of ML, with most of the difference coming from ML quiet Sun, whereas the sunspot areas are very similar. The discrepancy between the 3T facular and the HF unipolar areas is less than but qualitatively similar to that in the 1T

comparison, and the HF sunspot area is only about half that identified by the 3T method. Both the facular and sunspot areas for the 3T and ST methods agree better than for the 1T comparison, although roughly one-third of the ST active-region pixels are identified as either quiet-Sun or sunspot pixels by the 3T method.

Table 4 compares the ML labels with both the HF and ST classifications. The same combinations of categories are identified as quiet Sun for the HF and ST methods as in the previous two tables. Again, the HF unipolar area is much larger than the ML facular area, whereas the HF sunspot area is smaller. The comparison of ST active regions and ML faculae is similar to that for the ST – 1T case, whereas the ST and ML sunspot areas agree fairly well.

Finally, Table 5 compares the HF and ST labelings. To facilitate this comparison, we combine the HF weak-field and unassigned categories as best corresponding to ST quiet network, strong-field mixed polarity regions as ST enhanced network, and unipolar as ST active regions because the strong majority of ST active-region pixels correspond to HF unipolar pixels. However, this correspondence is not very good as the majority of HF unipolar pixels originate in ST quiet and enhanced network. The HF sunspot area is only about two-thirds the ST area, with most of the discrepant pixels being in the HF unassigned category. Many of these unassigned pixels are for large negative intensity contrasts with weak line-of-sight fields such as might occur in some areas of penumbra and near the limb where the field is nearly perpendicular to the line of sight.

## 4. Discussion

From the previous section one can see by visual inspection of the feature labels, from the confusion matrices, and from $\kappa$ that agreement is poorest when any of the methods are compared with HF. This is perhaps not surprising because the HF method was not designed to identify spatial features. For example, the total sunspot area is lower for the HF method than for any of the other algorithms. This is mostly because the histogram subdomains do not completely span the variable space. In particular, dark features that lack strong line-of-sight field, as will occur most often in penumbrae near the limb, and strong-field features that are in sunspots but are not especially dark are missed by the HF technique and are counted as unassigned. This is reflected both in the large number of unassigned pixels near the southeast spot in Figure 4e and in the summary tables. However, correlation of the histogram factors with total irradiance is comparable to that achieved with spatial features, and the analyses of Jones *et al.* (2000, 2003) imply that many features identified by histogram subdomains are temporally correlated with the features identified by the other techniques. The fact that $\kappa/\kappa_{\max}$ generally considerably exceeds $\kappa$ for HF comparisons also indicates that HF features either include or are included in features identified by other methods. Moreover, comparison of HF and ML feature areas over time (J.M. Pap, private communication) explicitly shows strong temporal correlation between ML sunspots and faculae with the corresponding HF features even though the actual areas differ substantially. Another important point of the HF technique is that the temporal variation of one magnetically well-determined factor (primarily strong mixed polarity fields) is not well correlated with irradiance variation.

Even where agreement between methods is comparatively good, there are still important differences, especially for faculae, but even in some instances for sunspots. Much of the disagreement for faculae results from their low contrast at the spatial resolution of the instruments producing the data for this paper. For this reason, many feature identification methods rely on higher contrast "proxy" measurements of faculae from Ca K (as in the CSUN observations), Mg k, or magnetograms (as in the ML, ST, and HF methods). The proxies are

usually formed higher in the solar atmosphere than faculae, tend to have larger but spatially correlated areas, and have center-to-limb visibility that is markedly different from faculae, which are best seen near the solar limb.

The methods also differ in their use of contiguity ("guilt by association") to determine class membership. The effect is on the spatial scale of the feature identifications. Solar physicists clearly differ in what they expect in this regard, and this may be largely a function of the purpose for which the identifications are intended. For example, a more spatially detailed classification scheme may be needed for irradiance comparison than is required for comparison with global dynamo models.

The comparisons between feature identification methods discussed here are far from complete but we feel that they are representative of methods in actual use. Given the diversity of observations, the variety of approaches, and lack of clear standards, improvement may be slow. However, we do suggest that there are areas where improvement is possible. Perhaps the most important is to establish agreement on what classes of features should be identified for a given purpose. If such agreement can be achieved, iterative development of standards becomes possible. Observations can be the basis for good physical models or physics-based feature definitions that allow prediction of feature appearance for different instruments under a variety of observing conditions. Iterative comparison between theory and observation then can in principle lead to more rigorous absolute standards against which various labeling algorithms can be tested. The beginnings of such development can be seen in the modeling of faculae by Keller *et al.* (2004) and the corresponding observations of Berger, Rouppe van der Voort, and Lofdahl (2005). Here, the link between accurate models and the spatial resolution of both observations and labeling algorithms becomes especially important since no instrument can presently image the full solar disk at the resolution of either the above model or observations.

It is perhaps worth noting that the very notion of features is dependent on human perception and is thus subjective. Presumably, some physical reality underlies the recognition of common image characteristics by educated observers, and all the methods attempt to reduce uncertainties introduced by differing individual perception by relying extensively on computer algorithms, which are in some sense objective. However, different rationales are used to set parameters. In the 1T and 3T methods, quantitative arguments are used to establish thresholds that minimize effects of seeing and confusion with instrumental noise, but these are dependent on instrument and observing site. The ML method requires an independent method for labeling features in a training set but does not otherwise depend on thresholds. The ST and HF methods use thresholds established by the knowledge and experience of the investigators in analyzing the SPM data. All are subjective in the selection of which features are of physical interest. A major goal of this paper is to begin testing the effects of these many choices by direct comparison.

**Acknowledgements** The authors are pleased to acknowledge P. Jones for help with quantitative comparison of categorical data and an anonymous referee for suggestions that substantially improved the original manuscript. H.P.J. carried out much of his early contribution to this research as a member of the Laboratory for Astronomy and Solar Physics at NASA's Goddard Space Flight Center (GSFC). This research was supported by several NSF and NASA Supporting Research and Technology and Guest Investigator grants over many years. SOHO is a mission of international cooperation between ESA and NASA, and the authors gratefully acknowledge the effort of the MDI team. NSO/KPVT data used here were produced cooperatively by AURA/NSO, NASA's GSFC, and NOAA/SEC.

## References


Berger, T.E., Rouppe van der Voort, L., Lofdahl, M.G.: 2005, High resolution magnetogram measurements of solar faculae. In: *AGU Spring Meeting Abstracts*, SP31 – A02.

Chapman, G.A., Cookson, A.M., Dobias, J.J.: 1996, Variations in total solar irradiance during solar cycle 22. *J. Geophys. Res.* **101**, 13541 – 13548. doi:10.1029/96JA00683.

Chapman, G.A., Cookson, A.M., Dobias, J.J.: 1997, Solar variability and the relation of facular to sunspot areas during solar cycle 22. *Astrophys. J.* **482**, 541 – 545. doi:10.1086/304138.

Chapman, G.A., Herzog, A.D., Lawrence, J.K., Walton, S.R., Hudson, H.S., Fisher, B.M.: 1992, Precise ground-based solar photometry and variations of total irradiance. *J. Geophys. Res.* **97**, 8211 – 8219.

Cohen, J.: 1960, A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37 – 46.

de Toma, G., White, O.R., Chapman, G.A., Walton, S.R., Preminger, D.G., Cookson, A.M., Harvey, K.L.: 2001, Differences in the Sun's radiative output in cycles 22 and 23. *Astrophys. J.* **549**, L131 – L134. doi:10.1086/319127.

de Toma, G., White, O.R., Chapman, G.A., Walton, S.R., Preminger, D.G., Cookson, A.M.: 2004, Solar cycle 23: an anomalous cycle? *Astrophys. J.* **609**, 1140 – 1152. doi:10.1086/421104.

Fligge, M., Solanki, S.K.: 2000, The solar spectral irradiance since 1700. *Geophys. Res. Lett.* **27**, 2157 – 2160. doi:10.1029/2000GL000067.

Fligge, M., Solanki, S.K., Unruh, Y.C.: 2000, Modelling irradiance variations from the surface distribution of the solar magnetic field. *Astron. Astrophys.* **353**, 380 – 388.

Foukal, P., Lean, J.: 1988, Magnetic modulation of solar luminosity by photospheric activity. *Astrophys. J.* **328**, 347 – 357. doi:10.1086/166297.

Foukal, P., Bernasconi, P., Eaton, H., Rust, D.: 2004, Broadband measurements of facular photometric contrast using the solar bolometric imager. *Astrophys. J.* **611**, L57 – L60. doi:10.1086/423787.

Fröhlich, C.: 2006, Solar irradiance variability since 1978. *Space Sci. Rev.* **125**, 53 – 65. doi:10.1007/s11214-006-9046-5.

Fröhlich, C., Lean, J.: 1998, The Sun's total irradiance: cycles, trends and related climate change uncertainties since 1976. *Geophys. Res. Lett.* **25**, 4377 – 4380. doi:10.1029/1998GL900157.

Harvey, K.L., White, O.R.: 1999, Magnetic and radiative variability of solar surface structures. I. Image decomposition and magnetic-intensity mapping. *Astrophys. J.* **515**, 812 – 831. doi:10.1086/307035.

Hoyt, D.V., Schatten, K.H.: 1993, A discussion of plausible solar irradiance variations, 1700 – 1992. *J. Geophys. Res.* **98**, 18895 – 18906.

Jones, H.P., Branston, D.D., Jones, P.B., Wills-Davey, M.J.: 2000, Analysis of NASA/NSO spectromagnetograph observations for comparison with solar irradiance variations. *Astrophys. J.* **529**, 1070 – 1083. doi:10.1086/308315.

Jones, H.P., Branston, D.D., Jones, P.B., Popescu, M.D.: 2003, Comparison of total solar irradiance with NASA/National Solar Observatory spectromagnetograph data in solar cycles 22 and 23. *Astrophys. J.* **589**, 658 – 664. doi:10.1086/374413.

Keller, C.U., Schüssler, M., Vögler, A., Zakharov, V.: 2004, On the origin of solar faculae. *Astrophys. J.* **607**, L59 – L62. doi:10.1086/421553.

Krivova, N.A., Solanki, S.K., Fligge, M., Unruh, Y.C.: 2003, Reconstruction of solar irradiance variations in cycle 23: is solar surface magnetism the cause? *Astron. Astrophys.* **399**, L1 – L4. doi:10.1051/0004-6361:20030029.

Lean, J.: 2001, The variability of the Sun: from the visible to the X-rays (CD-ROM directory: contribs/lean). In: Garcia Lopez, R.J., Rebolo, R., Zapaterio Osorio, M.R. (eds.) *11th Cambridge Workshop on Cool Stars, Stellar Systems and the Sun*, *ASP* **223**, Astron. Soc. Pac., San Francisco, 109 – 116.

Lean, J., Beer, J., Bradley, R.: 1995, Reconstruction of solar irradiance since 1610: implications for climate change. *Geophys. Res. Lett.* **22**, 3195 – 3198. doi:10.1029/95GL03093.

Lean, J., Rottman, G., Harder, J., Kopp, G.: 2005, SORCE contributions to new understanding of global change and solar variability. *Solar Phys.* **230**, 27 – 53. doi:10.1007/s11207-005-1527-2.

Pap, J.M.: 2003, Total solar and spectral irradiance variations from near-UV to infrared. In: Rozelot, J.P. (ed.) *The Sun's Surface and Subsurface: Investigating Shape*, *Lecture Notes in Physics* **599**, Springer, Berlin, 129 – 155.

Pap, J.M., Turmon, M., Floyd, L., Fröhlich, C., Wehrli, C.: 2002, Total solar and spectral irradiance variations from solar cycles 21 to 23. *Adv. Space Res.* **29**, 1923 – 1932.

Preminger, D.G., Walton, S.R., Chapman, G.A.: 2001, Solar feature identification using contrasts and contiguity. *Solar Phys.* **202**, 53 – 62.

Preminger, D.G., Walton, S.R., Chapman, G.A.: 2002, Photometric quantities for solar irradiance modeling. *J. Geophys. Res.* **107**, 1354 – 1362. doi:10.1029/2001JA009169.

Turmon, M., Pap, J.M., Mukhtar, S.: 2002, Statistical pattern recognition for labeling solar active regions: application to SOHO/MDI imagery. *Astrophys. J.* **568**, 396 – 407. doi:10.1086/338681.

Wenzler, T., Solanki, S.K., Krivova, N.A., Fröhlich, C.: 2006, Reconstruction of solar irradiance variations in cycles 21 – 23 based on surface magnetic fields. *Astron. Astrophys.* **460**, 583 – 595. doi:10.1051/0004-6361:20065752.

Willson, R.C., Hudson, H.S.: 1988, Solar luminosity variations in solar cycle 21. *Nature* **332**, 810 – 812. doi:10.1038/332810a0.

Willson, R.C., Mordvinov, A.V.: 2003, Secular total solar irradiance trend during solar cycles 21 – 23. *Geophys. Res. Lett.* **30**, 1199 – 1202.

Willson, R.C., Gulkis, S., Janssen, M., Hudson, H.S., Chapman, G.A.: 1981, Observations of solar irradiance variability. *Science* **211**, 700 – 702.