**Astronomy
&
Astrophysics**

The SPoCA-suite: Software for extraction, characterization, and tracking of active regions and coronal holes on EUV images

C. Verbeeck, V. Delouille, B. Mampaey, and R. De Visscher

STCE/Royal Observatory of Belgium, Avenue Circulaire 3, 1180 Uccle, Belgium
e-mail: cis.verbeeck@oma.be

ABSTRACT

Context. Precise localization and characterization of active regions (AR) and coronal holes (CH) as observed by extreme ultra violet (EUV) imagers are crucial for a wide range of solar and helio-physics studies.
Aims. We introduce a set of segmentation procedures (known as the SPoCA-suite) that allows one to retrieve AR and CH properties on EUV images taken from SOHO-EIT, STEREO-EUVI, PROBA2-SWAP, and SDO-AIA.
Methods. We build upon our previous work on the Spatial Possibilistic Clustering Algorithm (SPoCA), that we have improved substantially in several ways.
Results. We apply our algorithm on the synoptic EIT archive from 1997 to 2011 and decompose this dataset into regions that can clearly be identified as AR, quiet Sun, and CH. An antiphase between AR and CH filling factor is observed, as expected. The SPoCA-suite is next applied to datasets from EUVI, SWAP, and AIA. The time series pertaining to ARs or CHs are presented.
Conclusions. The SPoCA-suite enables the extraction of several long time series of AR and CH properties from the data files of EUV imagers and also allows tracking individual ARs or CHs over time. For AIA images, AR and CH catalogs are available in near-real time from the Heliophysics Events Knowledgebase. The full code, which allows processing any EUV images, is available upon request to the authors.

Key words. techniques: image processing – Sun: corona – Sun: activity – Sun: UV radiation

1. Introduction

Accurate determination of active region (AR) and coronal hole (CH) properties on coronal images is important for a wide range of applications. As regions of locally increased magnetic flux, the ARs are the main source of solar flares. For example a catalog of ARs describing key parameters, such as their location, shape, area, mean intensity, and integrated intensity, would allow us to relate those properties to the occurrence of flares. Having a bounding box of ARs can prove useful when studying several thousands of ARs together, for example, when performing a statistical analysis on oscillations of coronal loops.

Precise localization of CHs on the other hand is important because of the strong association between CHs and high-speed solar wind streams (Krieger et al. 1973). Finally, solar EUV flux plays a major role in solar-terrestrial relationships, and therefore, an accurate monitoring of AR, quiet Sun (QS), and CH is desirable as an input into solar EUV flux models.

In this paper, we present the SPoCA-suite, a set of algorithms that allows separation and extraction of AR, QS, and CH on EUV images. The SPoCA-suite includes the specific algorithms, Fuzzy C-means (FCM, Bezdek 1981), Possibilistic C-means (PCM, Krishnapuram & Keller 1993, 1996), PCM2, Spatial Possibilistic Clustering Algorithm (SPoCA), and SPoCA2, and histogram-based versions thereof. A detailed account is provided in Sect. 2. We also indicate how to select the parameters to optimize the segmentation of images taken by SOHO-EIT (Delaboudinière et al. 1995), STEREO-EUVI (Wuelser et al. 2004), PROBA2-SWAP (Berghmans et al. 2006; De Groof et al. 2008; Seaton et al. 2012), and SDO-AIA (Lemen et al. 2012). Combining these algorithms with pre- and postprocessing routines, the SPoCA-suite provides a powerful tool for consistent automatic detection of AR, QS, and CH, enabling systematic studies of their properties.

The SPoCA-suite serves as a module of the Feature Finding Team[1] (FFT; Martens et al. 2012), which is the main source of modules for the SDO Event Detection System (EDS; Hurlburt et al. 2012). While several EDS modules are run at the Smithsonian Astrophysical Observatory (with a lag time of a few days), SPoCA is one of the EDS modules that are run in near-real time at Lockheed Martin Solar and Astrophysics Laboratory. Every four hours, the EDS generates and uploads the SPoCA entries into the AR and CH catalogs of the HEK or Heliophysics Events Knowledgebase (Hurlburt et al. 2012).

The HEK is further linked through an API to the graphical interface isolsearch[2], the ontology software package of Solarsoft (SSW)[3], and the JHelioviewer visualization tool[4] (Müller et al. 2009).

Another option to access both solar images and metadata (including the HEK) through web user interfaces and the command line in IDL and Python is offered by the VSO or Virtual Solar Observatory (Hill et al. 2004; Gurman et al. 2012)[5]. The SPoCA-suite is written in C++ and contains wrappers in Python. The code is available upon request to the contact author.

---

[1] See http://solar.physics.montana.edu/sol_phys/fft/
[2] See http://www.lmsal.com/hek/hek_isolsearch.html for the HEK graphical interface and the link to the tutorial of the HEK.
[3] See http://sdoatsidc.oma.be/web/sdoatsidc/spoca_hek for explanation on how to use the SPoCA-suite within SolarSoft.
[4] See jhelioviewer.org
[5] See http://sdac.virtualsolar.org

The present paper is intended as a reference for the users of the SPoCA-suite, which builds upon our previous work on the Spatial Possibilistic Clustering Algorithm (Barra et al. 2009) and has seen major improvements. At its core lies a multichannel, unsupervised, fuzzy clustering method that segments EUV images into different regions according to their intensity level.

Development of automated solar feature detection and identification methods has increased dramatically in recent years due to the growing volume of data available. An overview of the fundamental image-processing techniques used in these algorithms is presented in Aschwanden (2010). These techniques are tailored to detect features in various types of observations at different heights in the solar atmosphere, as in Martens et al. (2012); Pérez-Suárez et al. (2011). For example, regions of locally intense magnetic flux are observed as dark spots (sunspots) in white light, CaII, or continuum images, as a concentration of strong positive and negative magnetic field values (magnetic AR) in magnetograms and as patches of enhanced brightness in both chromospheric imagery (plages) and coronal images (AR).

Image segmentation methods are typically classified into three broad categories: region-based methods, edge-based methods, and hybrid approaches. Region-based methods seek a partition of the image satisfying an homogeneity criterion on mono- or multispectral gray levels, or higher level attributes, such as texture or feature vectors that model pixels and their neighborhood. The dual edge-based approaches aim at characterizing image discontinuities and thus, locating region boundaries. Primal edge-based methods seek maximum intensity gradients, using either spatial filters, frequency filters, or zeros in the Laplacian of the image, and are often preprocessed by low pass Gaussian filtering, courtesy of the Laplacian's sensitivity to noise. Finally, the hybrid methods either consider a cooperation between region and contour approaches, or process some other original method.

Region-based methods for the detection of sunspots include thresholding against background (Pettauer & Brandt 1997; Colak & Qahwaji 2008), histogram-based thresholding (Steinegger et al. 1997), a region-growing method (Preminger et al. 1997), or a Bayesian approach (Turmon et al. 2002). Colak & Qahwaji (2008); Nguyen et al. (2005) combine thresholding and machine learning techniques to extract and classify sunspots according to the McIntosh system. Curto et al. (2008) and Watson et al. (2009) both use an edge-based approach based on mathematical morphology. Zharkov et al. (2005) uses an edge detection method combined with morphological operations, whereas Lefebvre & Rozelot (2004) present a singular spectrum analysis to detect sunspots and faculae at the solar limb.

The ARs, as observed by magnetograms, can be extracted and characterized by means of region growing techniques (Benkhalil 2003; McAteer et al. 2005; Higgins et al. 2011), thresholding in intensity (Qahwaji & Colak 2006; Colak & Qahwaji 2009), or wavelet domain (Kestener et al. 2010). Verbeeck et al. (2011) provide a detailed comparison of outputs from four automatic detection algorithms that detect sunspots, magnetic, and coronal ARs using six weeks of SOHO-EIT data. At the chromospheric level, network and plage regions are separated using thresholding methods (Steinegger et al. 1998; Worden et al. 1999), which are possibly combined with region-growing techniques (Benkhalil et al. 2006). Coronal ARs are segmented using either local thresholding, region-growing methods (Benkhalil et al. 2006), supervised techniques (Dudok de Wit 2006; Colak & Qahwaji 2013), or unsupervised techniques (Barra et al. 2009). Revathy et al. (2005) compares segmentation results of pixelwise fractal dimension of EIT images

using thresholding, region-growing techniques, and supervised classification.

Coronal holes are regions that have a lower electron density and temperature compared to the typical QS and thus, appear as dark regions in EUV and X-ray images. However, automated detection of CHs by intensity thresholding in one wavelength (for example, EIT 284 Å wavelength in Abramenko et al. 2009; Obridko et al. 2009; or soft X-ray images in Vršnak et al. 2007; Verbanac et al. 2011) is intrinsically complicated due to the presence of filaments and transient dimmings of a similar intensity level.

To resolve this ambiguity, it is necessary to make use of additional information coming from other wavelengths, from magnetograms, or from the time evolution of the feature as a means to check the consistency of a CH candidate with actual physical parameters. For example, Henney & Harvey (2005) first use a fixed thresholding based on the two-day average of He I 10 830 Å spectroheliograms and then check the unipolarity of the CH candidates using photospheric magnetograms. de Toma & Arge (2005) use a combination of fixed thresholdings on multiple wavelengths (the four SOHO-EIT bandpasses, He I 10 830 Å, magnetograms, and Hα images) to determine stringent criteria for a region that belongs to a CH, whereas de Toma (2011) uses a similar technique on synoptic maps. The approach in Scholl & Habbal (2008) is to first perform histogram equalization and fixed thresholding to extract low intensity regions on the four bandpasses of SOHO-EIT. In a second stage, statistics on magnetic field parameters measured by SOHO-MDI are evaluated to distinguish between filaments and CHs.

A similar methodology is used in Krista & Gallagher (2009); the difference is that it first detects low intensity regions using local histograms of SOHO-EIT, STEREO-EUVI, and Hinode-XRT images. Other methods include a watershed approach (Nieniewski 2002); perimeter tracing for polar CHs using morphological transform and thresholding (Kirk et al. 2009); and the use of imaging spectroscopy to separate QS from CH emission (Malanushenko & Jones 2005).

Finally, the classification approach of Dudok de Wit (2006); Colak & Qahwaji (2013); Barra et al. (2009) separates both ARs and CHs using brightness intensity as observed in one or multiple bandpasses. The output from other automated detection codes from the Feature Finding Team can be employed to remove filaments (AAFDCC, Bernasconi et al. 2005) from SPoCA's CH detections and to remove bright points (BP Finder, Saar & Farid 2011) from SPoCA's AR detections. Additionally, the polarity inversion line module (PIL Module, Martens et al. 2012; Jones 2004) can be used to separate out filaments again.

Here, we describe the SPoCA-suite and apply it to large datasets from four different EUV imagers. Several fundamental improvements have been made to the SPoCA-suite with respect to Barra et al. (2008) and Barra et al. (2009) in terms of robustness, accuracy, and stability. We have developed solutions for two fundamental problems that are native to the possibilistic techniques discussed in Sect. 2: coincident clustering and the stability of the η_i parameters. To enable automatic segmentation on a continuous stream of data from such sources as AIA, the smooth variation in center values has been accommodated. We have also introduced a much smoother limb brightness correction, yielding more accurate AR and CH segmentations near the solar limb. Dedicated routines have been written to take care of proper region extraction, tracking, region statistics, and the creation of masks and overlay images. Whereas the SPoCA-suite in Barra et al. (2008) and Barra et al. (2009) was applied to EIT

data only, we now have made a detailed survey of datasets from EIT, EUVI, SWAP, and AIA with a tailored algorithm for each.

The SPoCA-suite allows us to identify and characterize both individual ARs or CHs and the total class of all AR or CH pixels in an EUV image. Individual regions can also be tracked over time. Output products include masks, contours, contour overlays onto the original image, and movies. For any individual AR or CH, the SPoCA-suite provides the location of its barycenter and of the corners of a bounding box. For every individual region and every total class, intensity moment statistics and properties, such as area and filling factor, are computed.

The SPoCA-suite's automatic detection scheme enables the construction of large time series of properties that pertain to AR or CH, which would be too large a task to perform on a manual basis. Examples of derived AR and CH time series are presented for each instrument in Sect. 3.

As part of the Feature Finding Team, programs have been specifically written to run the SPoCA-suite in near-real time on SDO-AIA images. The two corresponding modules that are included in the SDO EDS are called SPoCA-AR and SPoCA-CH, for ARs and CHs respectively. The corresponding catalogs of ARs and CHs are included in the SSW/IDL software and as such are widely available to the solar physics community[6].

Section 2 describes the algorithms in the SPoCA-suite, including pre- and postprocessing, region extracting, and tracking of individual regions over time. In Sect. 3.1, the filling factors, median and total intensity of the AR and CH classes are extracted over more than one solar cycle, employing SOHO-EIT data from 1997 to 2011. The SPoCA-suite is applied to three months of STEREO-EUVI data in Sect. 3.2 and to nine months of PROBA2-SWAP data in Sect. 3.3. The evolution of twelve months of SDO-AIA AR, QS, and CH filling factors in 2011 is presented in Sect. 3.4.

A complete archive of several of these time series (updated in near-real time) is available in the Solar Timelines viewer for AFFects (STAFF[7]), a dedicated viewer for solar activity, solar wind, and geomagnetic timelines developed at the Royal Observatory of Belgium within the European FP7 project AFFECTS (Advanced Forecast for Ensuring Communications Through Space).

2. The SPoCA-suite

To segment an EUV image into AR, QS, and CH, we need to attribute every pixel to one of these three classes, based on their pixel values in the image. The pixels with highest intensity correspond to AR, those with intermediate pixel values correspond to QS, and the pixels with lowest pixel values correspond to CH. We can also combine the information present in p corresponding images in different EUV channels to obtain a segmentation of the corona into AR, QS, and CH. In this case, every pixel corresponds to exactly one p-dimensional vector of pixel values, which we call a feature vector. Due to the physical differences between AR, QS, and CH, it is expected that these classes naturally form clusters in the space of feature vectors. Hence, we need a mathematically sound way to classify the set of feature vectors into three clusters. In a more general setting, we can consider the number of classes to be C.

One of the most simple clustering techniques is called K-Means. This is an iterative method, where every pixel at each iteration is assigned to some class, and the center of each class is determined. The initialization step consists of selecting an arbitrary but preferably well-chosen center $b_i \in \mathbb{R}^p$ for every class i. Every subsequent iteration consists of two steps. First, every feature vector is attributed to some class, based on the distance of the feature vector to the class centers. Second, the class center for every class is updated (it is the average of all feature vectors belonging to that class). These two steps are repeated until the subsequent class center values converge. K-Means is what we call a crisp clustering technique.

A more general clustering scheme is offered by fuzzy clustering. In this approach, a pixel is not assigned to exactly one class, but every pixel obtains a membership value to each of the classes. Similar to the K-Means method, the initialization step consists of selecting a center $b_i \in \mathbb{R}^p$ for every class i, and every subsequent iteration consists of two steps: First, the membership value of every feature vector to each class is calculated; second, the class center for every class is updated. Again, these steps are repeated until convergence of the class centers sets in with the result being a final membership value of every feature vector to every class. We can then apply some scheme to assign every feature vector to exactly one class, such as the class for which its membership value is highest. Figure 1 provides an illustration of several steps in this iterative process. Note that the pixel coordinates are not considered in either of the above methods.

The SPoCA-suite implements three types of fuzzy clustering algorithms that are specially tailored to the segmentation of solar coronal EUV images: the Fuzzy C-means (FCM); a regularized version of FCM known as Possibilistic C-means (PCM) algorithm, and a Spatial Possibilistic Clustering Algorithm (SPoCA) that integrates neighboring intensity values.

The description of the segmentation process in terms of fuzzy logic was motivated by the facts that information provided by an EUV solar image is noisy and subject to both observational biases (line-of-sight integration of a transparent volume) and interpretation (the apparent boundary between regions is a matter of convention). Fuzzy measures are able to represent ill-defined classes (without a clear-cut boundary) in a natural way. Furthermore, fuzzy segmentation methods more often reach a global optimum, rather than merely a local optimum, as compared to crisp clustering methods (Trauwaert et al. 1991). For more details on data clustering, we refer to Gan et al. (2007).

The mathematical description below helps us introduce the two most used algorithms of the SPoCA-suite. Let N be the number of pixels in each image, and let $\boldsymbol{x}_j \in \mathbb{R}^p$ be a p-dimensional feature vector that describes the Sun at a particular location. In our case, $\boldsymbol{x}_j$ is a p−dimensional vector corresponding to the intensities recorded at pixel j in p different channels. A fuzzy clustering algorithm searches for C different compact clusters among the $\boldsymbol{x}_j$'s in the set $X = \{\boldsymbol{x}_j \in \mathbb{R}^{\mathbf{p}} | 1 \leq j \leq N\}$ of all feature vectors. It does so by computing both a fuzzy partition matrix $\mathbf{U} = (u_{ij})_{1 \leq i \leq C, 1 \leq j \leq N}$ and the cluster centers $B = \{b_i \in \mathbb{R}^p | 1 \leq i \leq C\}$. The scalar $u_{ij} = u_i(\boldsymbol{x}_j) \in [0, 1]$ is called the membership value of feature vector $\boldsymbol{x}_j$ to class i (Bezdek 1981).

2.1. Fuzzy C-Means algorithm (FCM)

Since its introduction by Bezdek (1981), the Fuzzy C-Means (FCM) algorithm has been widely used in pattern recognition and image segmentation in various fields, which include medical imaging (Philipps et al. 1995; Bezdek et al. 1997), remote
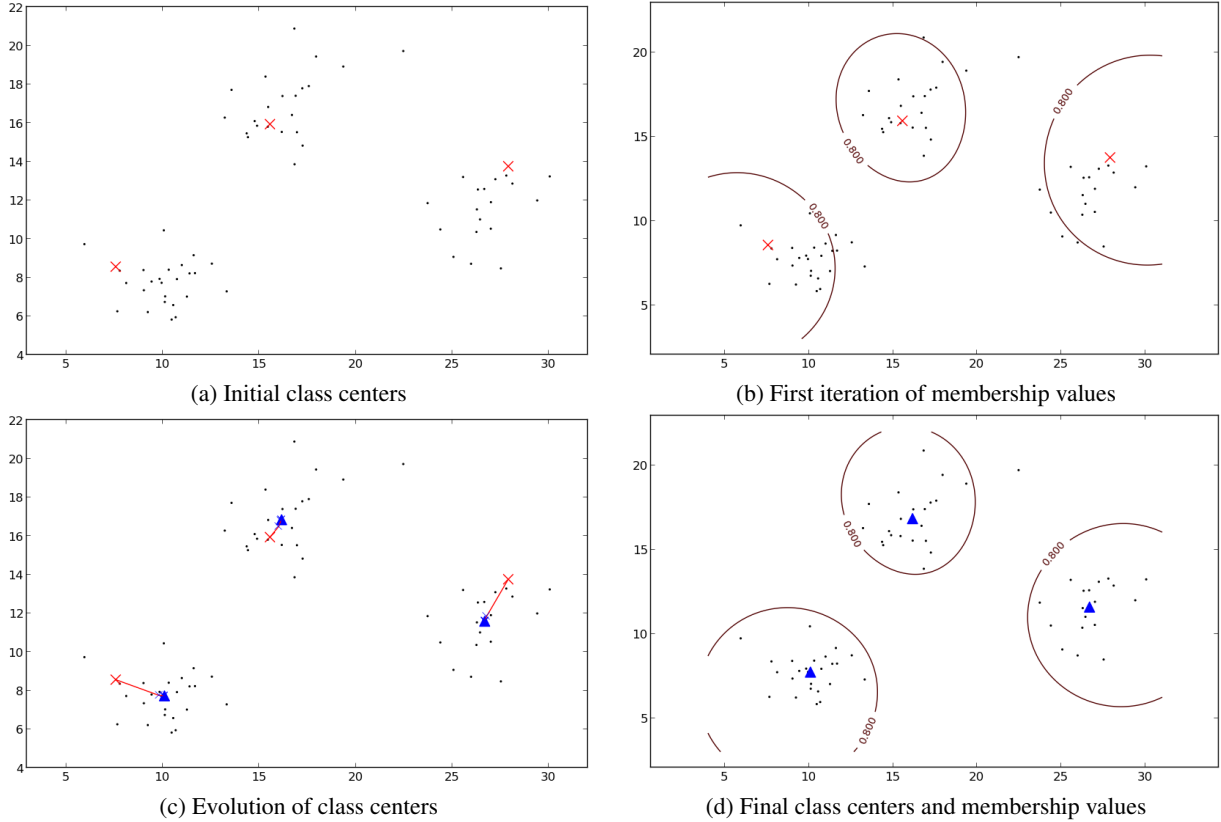
(a) Initial class centers

(b) First iteration of membership values

(c) Evolution of class centers

(d) Final class centers and membership values

Fig. 1. Illustration of several steps in the iterative procedure of FCM on a simple synthetic 2D dataset containing three clusters of data points. **a)** An initial center (red cross) is chosen for every class. **b)** In a first iteration, the membership values for every data point are calculated, considering the class centers. The contours delineate a membership value of 0.8. **c)** In a next step, the new class centers are calculated, considering the new membership values. After seven such two-step iterations, the class centers have converged to their final values, indicated by the blue triangles. **d)** In the last step, the final membership values are calculated, considering the final class centers. Note that both the final position of the centers and the final membership contours correspond much better to the three clusters than the initial ones.

sensing imaging (Rangsanseri & Thitimajshima 1998; Melgani et al. 2000), and image segmentation in vision (Baker et al. 2003).

The idea behind both K-Means and FCM is the minimization of the total crisp or fuzzy intracluster variance. In the crisp case, the intracluster variance of class i is the sum of squared distances between every point of class i and the class center b_i. The fuzzy intraclass variance of class i is defined as $\sum_{j=1}^{N} u_{ij}^m d^2(x_j, b_i)$, where d is a metric in $\mathbb{R}^p$ (typically, the Euclidean distance). Hence we try to minimize the total fuzzy intracluster variance in FCM:

$$J_{\mathrm{FCM}}(B, U, X) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d^2(x_j, b_i), \qquad (1)$$

which is subject to

$$\forall i \in \{1, \cdots, C\} : \sum_{j=1}^{N} u_{ij} < N \text{ and } \forall j \in \{1, \cdots, N\} : \sum_{i=1}^{C} u_{ij} = 1, \qquad (2)$$

where m is a parameter that controls the degree of fuzzification ($m = 1$ means no fuzziness). In practice, a value of $m = 2$ is often chosen, as it allows for a fast computation in the iterative scheme. In our application, we consider either one ($p = 1$) or two channels ($p = 2$).

The minimization of (1) is reached when

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d^2(x_j, b_i)}{d^2(x_j, b_k)} \right)^{1/(m-1)}} \text{ and} \qquad (3)$$

$$b_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}. \qquad (4)$$

Formulas (3) and (4) are used in steps 1 and 2, respectively, of every iteration and were obtained by gradient descent of $J_{\mathrm{FCM}}(B, U, X)$.

There are two types of shortcomings of FCM. First, it is sensitive to noise and outliers (Krishnapuram & Keller 1993). Second, it is theoretically not satisfying, since the membership degree of a feature vector with respect to any class in (3) depends on its distances to all other class centers.

Following our study, FCM yields the best results for extracting CHs out of the (almost noise free) 193 Å AIA images. Given the large size of AIA images, however, FCM is applied on the histogram of normalized intensity values, rather than on individual values. In this approach, each feature vector is rounded to the center of the histogram bin to which it belongs. In this way, the number of feature vectors is reduced to the number of histogram bins, which considerably reduces computation time. The normalization consists in dividing by exposure time, correcting for limb brightness enhancement, and dividing by the median value (see Sect. 2.5). A bin size of 0.01 DN/s for the histogram provides the same precision as when individual values are used.

2.2. Possibilistic C-means algorithm (PCM)

To obtain a formulation for u_{ij} that depends only on the distance of feature vector x_j to the center of class i, Krishnapuram & Keller (1993, 1996) proposed the minimization of the objective function

$$J_{\text{PCM}}(B, U, X) = \sum_{i=1}^{C} \left(\sum_{j=1}^{N} u_{ij}^m d^2(x_j, b_i) + \eta_i \sum_{j=1}^{N} (1 - u_{ij})^m \right), \quad (5)$$

which is subject to

$$\forall i \in \{1, \cdots, C\} : \sum_{j=1}^{N} u_{ij} < N \text{ and } \forall j \in \{1, \cdots, N\} : \max_i u_{ij} > 0. \quad (6)$$

The first term of J_{PCM} in (5) is the intracluster variance, whereas the second term enforces u_{ij} to depend only on $d(x_j, b_i)$ and stems from the relaxation of the probabilistic constraint in (2).

Parameter η_i in (5) is homogeneous to a squared distance. It can be fixed, or updated at each iteration. Krishnapuram & Keller (1993) proposed the computation of η_i as the intraclass dispersion:

$$\eta_i = \frac{\sum_{j=1}^{N} u_{ij}^m d^2(x_j, b_i)}{\sum_{j=1}^{N} u_{ij}^m}. \quad (7)$$

The solution of the minimization of (5) satisfies

$$u_{ij} = \left[1 + \left(d^2(x_j, b_i)/\eta_i \right)^{\frac{1}{m-1}} \right]^{-1} \quad \text{and} \quad (8)$$

$$b_i = \sum_{j=1}^{N} u_{ij}^m x_j / \sum_{j=1}^{N} u_{ij}^m. \quad (9)$$

In practice, PCM is initialized by a run of FCM, which allows for computation of η_i as in (7). Krishnapuram & Keller (1993) proved the convergence of iteration (8)−(9) for fixed values of η_i. PCM is more robust to noise and outliers than FCM and provides independent functions $\boldsymbol{u}_i = \{u_{ij} | j = 1, \ldots, N\}$. It must be corrected, however, from coincident clustering (Sect. 2.2.1), and a proper choice of the parameter η_i must be made (Sect. 2.2.2). The SPoCA-AR module of the HEK uses this corrected PCM algorithm on the AIA 171 Å and 193 Å bandpasses. Similar to the SPoCA-CH module, it is applied on histogram intensity values rather than on individual pixel intensity values.

The Spatial Possibilistic Clustering Algorithm (SPoCA) that integrates neighboring intensity values was described in detail in Barra et al. (2008, 2009). It is basically a version of PCM, where each contribution of a pixel x_j in formulas (5), and (7)−(9) has been replaced by the weighted average contribution of all pixels in a small spatial neighborhood of x_j.

2.2.1. Coincident clustering

The original PCM suffers from convergence to a unique center, where there is often only one or two distinct clusters detected instead of three. This is a typical feature of possibilistic clustering algorithms called coincident clustering (Krishnapuram & Keller 1996). To circumvent this problem, we use special membership

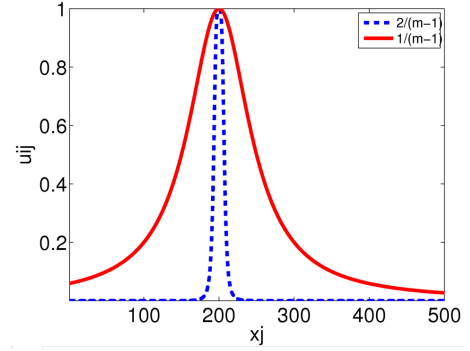**Fig. 2.** Comparison of membership functions u_{ij} for $b_i = 200$, when the exponent is chosen equal to $1/(m − 1)$ (red line) and $2/(m − 1)$ (blue dashed line). The blue dashed line is more compact, which leads to distinct class centers.

functions u_{ij}, which are more compact and hence, do not overlap so easily. More precisely, exponent $1/(m − 1)$ in (8) is replaced by $2/(m − 1)$; see Fig. 2 for a graphical representation. We name PCM2 the algorithm where the exponent in the membership function u_{ij} is equal to $2/(m − 1)$. The modified SPoCA algorithm where the exponent in the membership function u_{ij} is taken equal to $2/(m − 1)$, is called SPoCA2.

2.2.2. Constraints on the parameter η_i

The dynamical range of intensities differs among the AR, QS, and CH classes. In particular, the ARs show the largest spread in intensities. The parameter η_i, as computed in (7), can be viewed as a measure of dispersion or variance within a class. In the case that η_{AR} becomes prohibitively large, the value of $u_{\text{AR},j}$ for dark pixels x_j can be higher than $u_{\text{QS},j}$ or $u_{\text{CH},j}$, as illustrated in Fig. 3a. To avoid this situation, we enforce the following inequalities, as derived in Appendix A:

$$\frac{\eta_{\text{QS}}}{\eta_{\text{CH}}} < \frac{b_{\text{QS},q}}{b_{\text{CH},q}}, \frac{\eta_{\text{AR}}}{\eta_{\text{CH}}} < \frac{b_{\text{AR},q}}{b_{\text{CH},q}}, \frac{\eta_{\text{AR}}}{\eta_{\text{QS}}} < \frac{b_{\text{AR},q}}{b_{\text{QS},q}} \text{ for } q = 1, \ldots, p, \quad (10)$$

with $b_{\text{AR},q}, b_{\text{QS},q}$, and $b_{\text{CH},q}$ as the values for the qth channel of the class centers for AR, QS, and CH respectively. Figure 3 shows an example of a segmentation with and without constraints on η_i. Without constraints, some CH areas get classified in the same class as the AR. This problem is solved, when constraints (10) are introduced.

For solar EUV images, the combination of the iteration schemes (8), (9), and (7) tends to produce η_{CH}-values that converge to zero. Due to the condition (10) on η_{QS} and η_{AR}, these two parameters are also dragged along to converge to zero. Our iterative scheme therefore freezes the value of η_i, when it has changed by a factor α with respect to its starting value. In other words, formula (7) is used until iteration it, where

$$\eta_i^{(it)}/\eta_i^{(1)} > \alpha \text{ or } \eta_i^{(1)}/\eta_i^{(it)} > \alpha.$$

For the next iterations, we keep η_i^{it}. Satisfactory results have been obtained on a variety of datasets and instruments with $\alpha = 100$.

2.3. Smooth variation in center values

To have a smooth variation in the center values over time, the centers chosen at time t in the HEK are the median of the last 10 centers obtained at previous time stamps $t, t − 1, \ldots, t − 9$. To get the membership map corresponding to this smoothed value
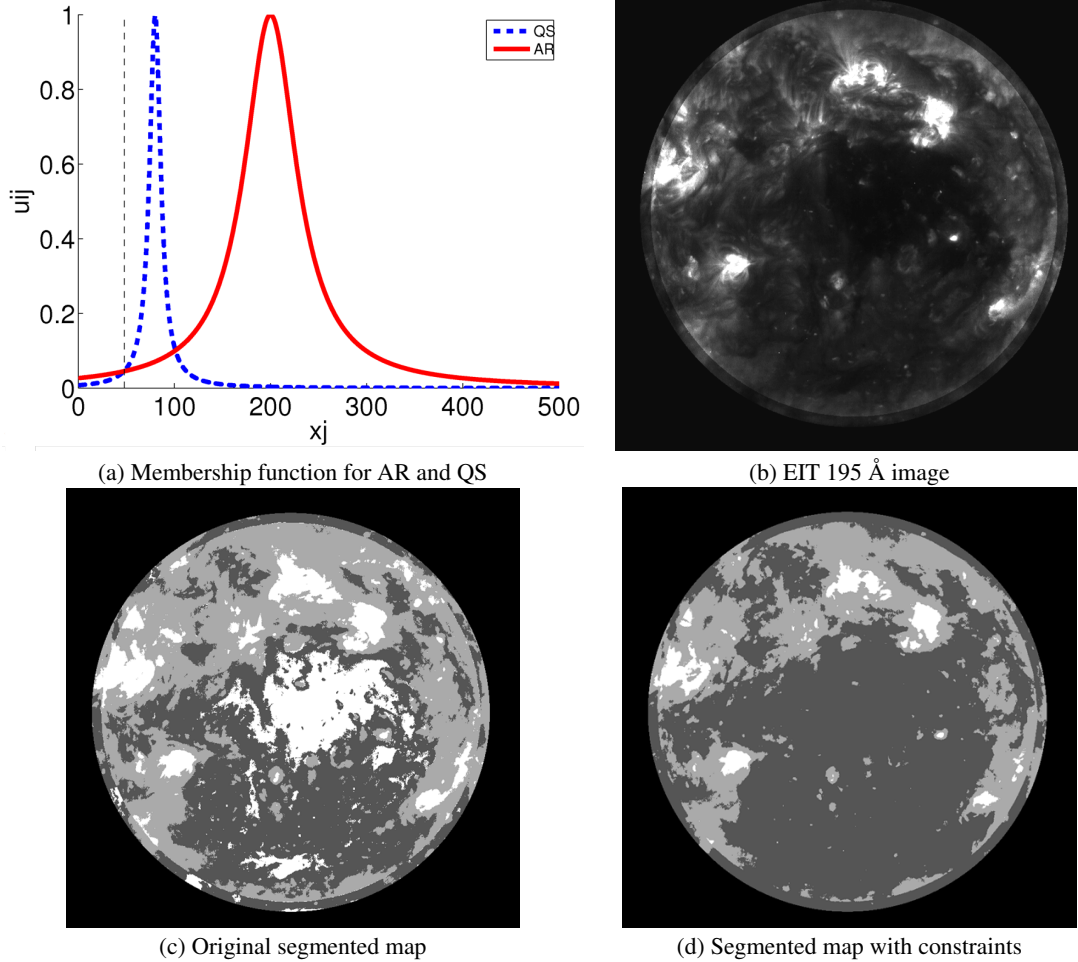
(a) Membership function for AR and QS

(b) EIT 195 Å image

(c) Original segmented map

(d) Segmented map with constraints

Fig. 3. a) Illustration of membership functions u_{ij} for AR and QS feature vectors x_j in the case of a large value of η_{AR}. Because of the larger spread in intensity values of the AR, small values of x_j may have a larger AR membership than QS membership. **b)** EIT 195 Å image from January 1, 2000 with limb brightness correction. **c)** Segmentation using original PCM algorithm: The darkest parts are classified as active regions. **d)** Segmentation using PCM with constraints on η_i-values.

of the center b_i, an attribution procedure using (3) or (8) is performed. This means that the final segmentation is obtained by applying (3) or (8) only once (using the smoothed center values) instead of using iterations until convergence.

2.4. Segmented maps

Given the membership maps U and centers B, a segmented map can be obtained using various decision rules.

Maximum. The most common rule is to attribute a pixel j to the class c for which it has the maximum membership value: $u_{cj} = \max_{i \in \{1,...,C\}} u_{ij}$. This rule is used in the SPoCA-CH module of the SDO EDS.

Threshold. On EUV images, the QS class contains typically the largest number of pixels, and hence the center of the QS class is the most stable over time. In contrast, the cardinality of points belonging to the AR class varies a lot over time, resulting in a high variation in the center of the AR class. To have a stable segmentation of AR over time, the SPoCA-AR module of the SDO EDS, therefore, uses a threshold on the QS membership class to decide whether a pixel belongs to the AR class. For the SPoCA-AR module, all pixels j whose values are higher than b_{QS} and for which $u_{QS,j}$ is lower than 0.0001 are attributed to the AR class.

Closest. This rule attributes a pixel to the class for which the Euclidean distance to the class center is the smallest.

Merge. A more complex procedure using sursegmentation and merging of classes has been described by Barra et al. (2009). Sursegmentation consists of segmenting the image into a number of classes strictly superior to the intuitively expected number of classes in the image and then finding an aggregation criterion of the resulting partition to exhibit the relevant classes.

Fix. This scheme applies the rule for the maximum as above and then merges the C resulting classes according to a fixed scheme. For instance, we can merge the inital $C = 4$ classes into three classes (AR, QS, CH) using CH = {1}, QS = {2, 3}, and AR = {4}. This rule is used in the analysis of the EIT archive in Sect. 3.1.

2.5. Preprocessing

Some preprocessing steps are needed to obtain an accurate segmentation of EUV images.

First, images can be calibrated using Solarsoft [8] routines, and intensities can be normalized by their median values. Second, the limb brightening effect observed in solar EUV images should

---

[8] http://www.lmsal.com/solarsoft/

Table 1. Values of parameters, r_1, r_2, r_3, and r_4 used for the limb brightness correction formula, expressed as percentage of the solar radius.

Instrument	r_1	r_2	r_3	r_4
EIT	80	95	107	112
EUVI	70	95	108	112
SWAP	95	100	105	115
AIA	70	95	108	112

be corrected before any segmentation based on intensity can be reliably applied. Finally, the SPoCA-suite can be applied on either linear or square-root transformed images. For Poisson data, Anscombe (1948) showed that a square-root transform induces exact asymptotic normality and stabilizes the variance. This is especially useful for the extraction of low-intensity regions, such as CHs, which are affected by Poisson noise.

A first limb brightening correction has been proposed by Barra et al. (2009). It consists of applying a polar transform to represent the image I in a (ρ, θ) plane with the origin at the solar disk center. The polar transform is a conformal mapping from points in the Cartesian plane (x, y) to points in this polar plane, as described by: $\rho = \sqrt{x^2 + y^2}, \theta = \text{atan}(y/x)$. We then computed the integral $F(\rho) = \frac{1}{2\pi}\int_0^{2\pi} I(\rho, \theta)\mathrm{d}\theta$, which specifies the intensity distribution as a function of ρ.

Denoting $m_\odot$ as the median value of intensities on the on-disk part of the corona, the image I_{corr} corrected for the enhanced brightness near the limb is computed as

$$I_{\text{corr}}(\rho, \theta) = m_\odot \frac{I(\rho, \theta)}{F(\rho)} \tag{11}$$

for values of ρ ranging between $0.95\,R_\odot$ and $1.05\,R_\odot$. Finally, I_{corr} is remapped to the Cartesian plane.

This abrupt correction leads to discontinuities in the images around these radial distances, as can be seen on the images in Fig. 3. We instead propose to apply a correction I_{smooth}:

$$I_{\text{smooth}}(x, y) = (1 - f(\rho(x, y)))\, I(x, y) + f(\rho(x, y))I_{\text{corr}}(x, y), \tag{12}$$

where $f(\rho(x, y))$ introduces a smooth transition between the zones not corrected (when ρ is between 0 and r_1 or above r_4) and the zones $\rho \in [r_2, r_3]$ that are fully corrected as described by Eq. (11). The transition function $f(\rho)$ is defined as

$$f(\rho) = \begin{cases} 0 & \text{if } \rho \leq r_1 \text{ or } \rho \geq r_4 \\ 1 & \text{if } \rho \in [r_2, r_3] \\ \frac{1}{2}\sin\left(\frac{\pi}{r_2 - r_1}(\rho - \frac{r_1 + r_2}{2})\right) + \frac{1}{2} & \text{if } r_1 \leq \rho \leq r_2 \\ \frac{1}{2}\sin\left(\frac{\pi}{r_4 - r_3}(\rho + \frac{r_4 - 3r_3}{2})\right) + \frac{1}{2} & \text{if } r_3 \leq \rho \leq r_4. \end{cases} \tag{13}$$

A parameter study has been performed to determine the optimal values of r_1, r_2, r_3, and r_4 for EIT, EUVI, SWAP, and AIA, see Table 1.

2.6. Region extraction and postprocessing

To extract individual regions (that are labeled as ARs or CHs) from segmented maps, the following postprocessing steps are implemented:

1. Compute a sinusoidal projection map (Snyder 1987). This improves the determination of regions toward the limb.
2. Clean the segmented map by removing elements smaller than 6 arcsec using a morphological erosion.

3. Aggregate neighboring blobs by performing a morphological closing, which consists of a dilation by 32 arcsec followed by an erosion.
4. Compute the inverse of the sinusoidal projection.

The reader is referred to (Barra et al. 2009) for an introduction to mathematical morphology in this context. The equirectangular and Lambert cylindrical projections are also implemented in the SPoCA-suite. Our test on SDO-AIA data shows that the sinusoidal projection provides the best results.

To remove bright points, a final cleaning can be performed as follows. The ARs smaller than 1500 arcsec^2 and the CHs smaller than 3000 arcsec^2 are discarded. Except for AIA, note that neither of the steps above was performed in the studies in Sect. 3, since the AR, QS, and CH classes are investigated as a whole, and hence, no individual regions were extracted.

A typical CH possesses relatively smooth boundaries. Having an accurate estimation of its shape and localization is important for space weather purposes, since CHs that are located at the central meridian and near the equator provoke the most intense geoeffective conditions (a few days later). We have computed a chain code for the CHs using a maximum of 100 points by default to describe the contour. The details of the algorithm are described in Appendix B.

The SPoCA-suite offers the option to either extract regions on the whole image, the solar disk (using a radius $r = R_\odot$), or another disk centered on the center of the Sun (e.g., choosing a radius $r = 1.3\,R_\odot$). It also allows us to identify and characterize both individual AR or CH and the total class of all AR or CH pixels in an EUV image. SPoCA-suite output products include (FITS) masks and contours, contour overlays onto the original image, and overlay movies.

For any individual AR or CH, the SPoCA-suite provides the Stonyhurst and Carrington longitude and latitude values corresponding to its barycenter and the corners of a bounding box. For every individual region and every total class, it also calculates the raw and deprojected area, filling factor, and statistics about the distribution of its pixel vales (minimum, maximum, median, average, variance, skewness, kurtosis, and total intensity)[9]. Hence, the SPoCA-suite's automatic detection scheme enables the construction of large time series of properties pertaining to AR or CH, which would be too large a task to perform on a manual basis. Examples of derived AR and CH time series are presented for each instrument in Sect. 3.

2.7. Tracking over time

Individual regions can also be tracked over time. Following the aggregation of regions described in the previous section, an AR is defined as a coherent group of corresponding AR blobs, and a CH is defined as a coherent group of corresponding CH blobs. The goal of tracking is to appoint the same ID number to a physical region (AR or CH) over time. A region observed at time stamp t can correspond to a single region at the next time stamp, but it can also split (and produce two children), or it can merge with neighboring regions.

Our tracking scheme amounts to creating a directed graph (N, E), where N is the list of nodes representing individual regions and E is the list of edges between regions. An edge between a region observed at time t_1 and another observed at time t_2 is created if their time difference $t_2 - t_1$ is smaller than

---

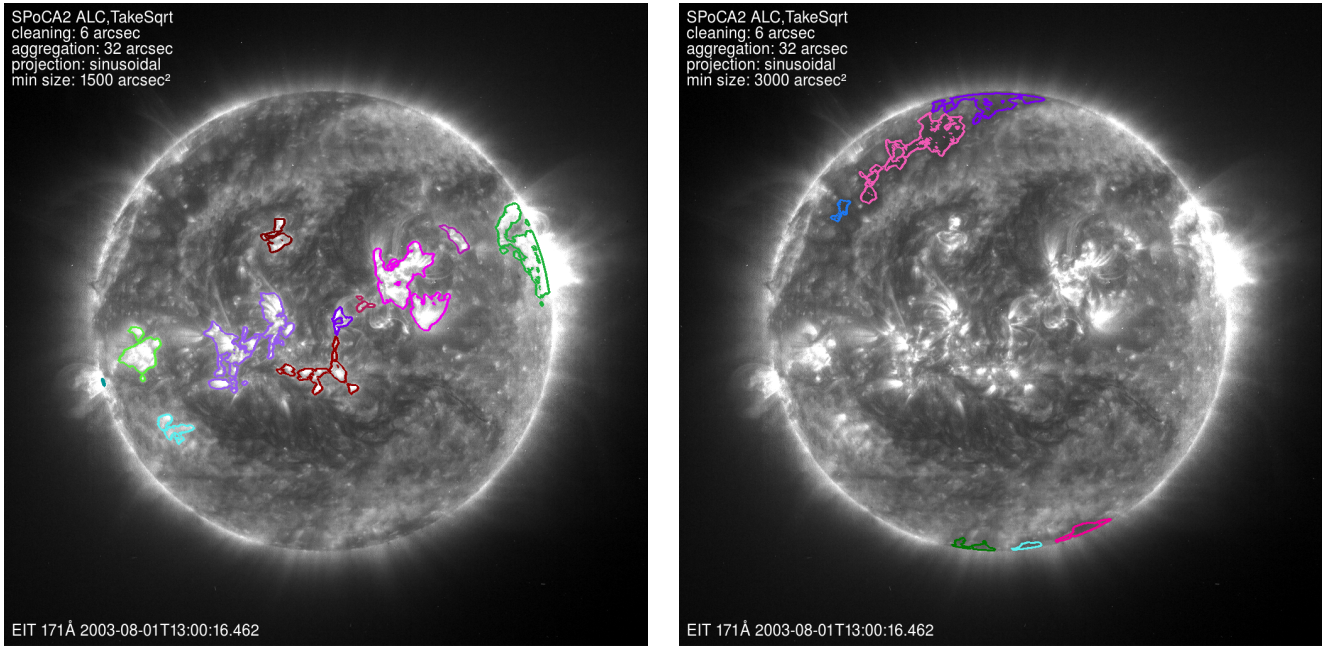[9] The list of all features computed can be found in http://www.lmsal.com/hek/VOEvent_Spec.html

Fig. 4. Overlay of **a)** (*left*) an AR map and **b)** (*right*) a CH map, created using EIT 171 and 195 Å, onto the corresponding EIT 171 Å image taken on August 1, 2003 around 13ʰ UT.

Table 2. Selected SPoCA-suite parameter values for the detection of active regions.

Active regions	EIT	EUVI	SWAP	AIA
Image transformation	sqrt	sqrt	linear	linear
Bandpasses used (Å)	(171, 195)	195	174	(193,171)
Number of classes C	4	4	6	4
Classifier	SPoCA2	SPoCA2	SPoCA2	HPCM2
Center values	avg(fix 42)	med(all)	raw	med(last 10)
Segmentation	Fix	Max	Fix	Thresh

Table 3. Selected SPoCA-suite parameter values for the detection of coronal holes.

Coronal holes	EIT	EUVI	AIA
Image transformation	sqrt	sqrt	sqrt
Bandpasses used (Å)	(171, 195)	195	193
Number of classes C	4	4	4
Classifier	SPoCA2	SPoCA2	HFCM
Center values	avg(fix 42)	med(all)	med(last 10)
Segmentation	Fix	Max	Max

some value, if they overlap, and if there is not already a path between them. It is possible to first derotate the region maps before comparing them. This is necessary, when large time differences are involved.

CHs are long-lived. Preliminary analysis shows that a CH candidate detected for more than three consecutive days exhibits the expected magnetic properties characteristic of unipolar regions. Hence, we include this temporal information in the tracking and report it to the HEK only for those CHs that are older than three days.

3. Results

In this paper, we are mainly looking for AR and CH features, and we leave out QS as the complement of the AR and CH classes. The SPOCA-suite allows us, however, to treat QS as a category (or class) in itself. A parameter study has been conducted to determine the optimal values of SPoCA-suite parameters for segmenting and extracting AR and CH properties on EIT, EUVI, SWAP, and AIA images. Tables 2 and 3 present a summary of the main parameters used in the extraction of AR and CH, respectively. Though these dedicated parameters provide well-chosen AR and CH detections on the imagers listed below, it should be noted that it is impossible to have an exact correspondence between the AR or CH detected by two different imagers.

3.1. Analysis of the synoptic SOHO-EIT archive

The Extreme Ultraviolet Telescope EIT (Delaboudinière et al. 1995) onboard SOHO delivers synoptic observations that consist of $1k \times 1k$ images of the solar corona recorded in four different wavelengths every six hours. The relatively small size of the SOHO-EIT images allows fast computations. The SOHO spacecraft however is situated at the L1-Lagrange point and is thereby exposed to cosmic ray hits (CRH) and to proton events, which deteriorate the quality of the image. The spatial regularization scheme SPoCA (or its improved version SPoCA2) proposed by Barra et al. (2009) is thus necessary to treat EIT images. Note that the SDO mission, on the other hand, operates in geosynchronous orbit. The SDO-AIA images are therefore less contaminated by CRH and suffer less from proton events.

Figure 4 shows an example of overlays of AR and CH maps onto the corresponding EIT images. Every color corresponds to a different AR or CH as identified by the SPoCA-suite's region extraction algorithm.

For the present study, we have downloaded all available, complete synoptic EIT FITS files (i.e., no missing blocks or other artifacts) from March 1, 1997 to August 17, 2011 and preprocessed them using the most current version of eit_prep in the Solarsoft library to obtain absolutely calibrated level 1 FITS files. Note that until recently, absolute calibration was only available for EIT FITS files dating from before 2006.
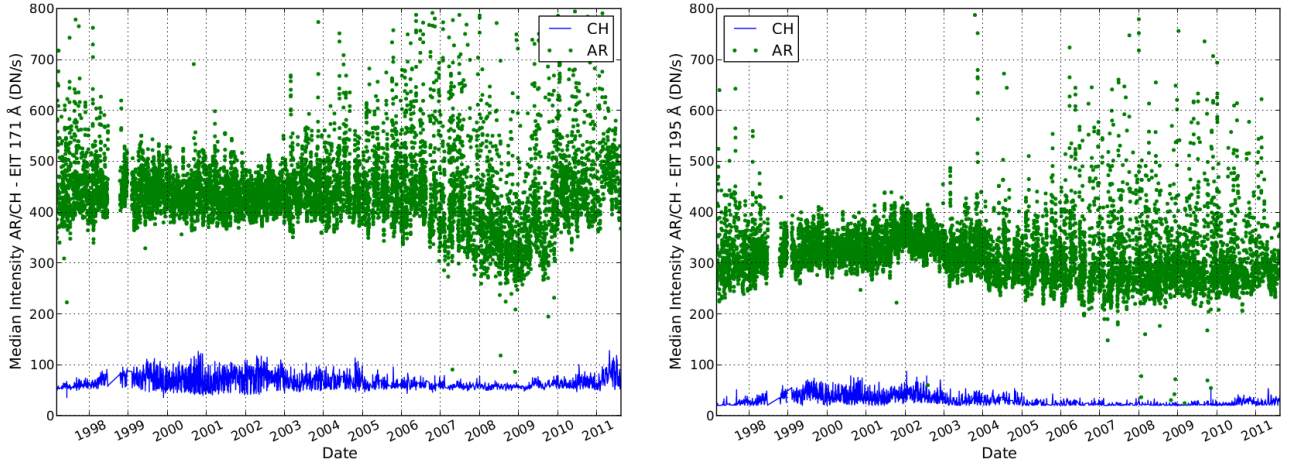
Fig. 5. Median intensity of AR and CH pixels for **a)** (*left*) EIT 171 Å and **b)** (*right*) EIT 195 Å from March 1, 1997 until August 17, 2011.
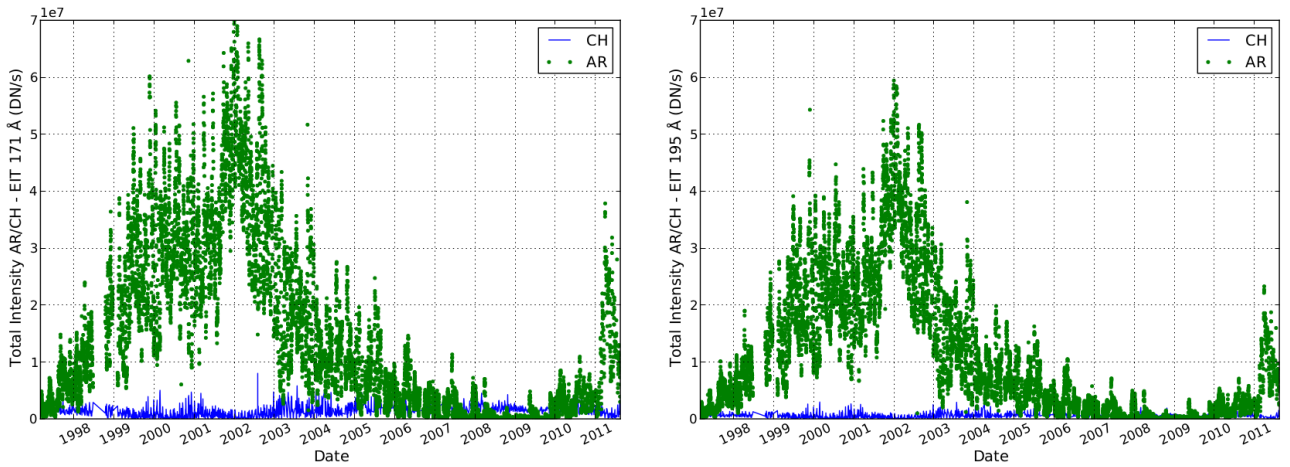

Fig. 6. Total intensity of AR and CH pixels for **a)** (*left*) EIT 171 Å and **b)** (*right*) EIT 195 Å from March 1, 1997 until August 17, 2011.

For the entire dataset, we used the combined data from 171 Å and 195 Å FITS files as input for SPoCA2. In a first stage, a segmentation into four classes was run with the square root of the pixels in the FITS files as input. The first class corresponds to CH, whereas the fourth class corresponds to AR. The overall quality of the segmentations was checked and acknowledged by human supervision on a representative subset of 112 pairs of FITS files.

Employing the average center and η_i values obtained during solar maximum (2000−2004, 42 pairs of FITS files) allowed us to perform a fixed-center SPoCA2 attribution on the whole dataset with the advantage that AR, QS, and CH are detected in a consistent way over the entire solar cycle, even in a low minimum when there are virtually no AR around. For instance, an AR or CH detected near the solar minimum would be detected in identical fashion if it appeared near a solar maximum when the whole image is brighter. The overall quality of these final segmentations was again checked and acknowledged on the subset of 112 pairs of FITS files. Statistics about pixel intensities and areas for AR, QS, and CH were computed for the final segmentations on the whole dataset, which yielded several time series with each one of them containing over 9500 data points. We present some of these time series below.

To study the evolution of AR, QS, or CH properties over the solar cycle, we can investigate time series of the median intensity, the total intensity, and the raw area or filling factor of the region. The raw area of a region is the area as measured in the plane of the image, as expressed in Mm^2. The filling factor of a region is defined as the raw area of this region in units such that the filling factor of the total solar disk equals 1. These filling factors can be included into (semi-) empirical models of the solar atmosphere, which can be employed to model the solar EUV irradiance (Haberreiter 2012). The total intensity of a region is the sum of all pixel values of the pixels inside that region. If the total intensity of the AR increases over time, for instance, it can be either due to an increase in the area of the AR or due to an inherent brightness variation in the loops in the AR. A measure of the inherent brightness of the plasma in AR, QS, and CH is provided by the median intensity of pixels within the region (which is a more robust version of the average intensity).

Figure 5 presents the median intensity (in DN/s) of all pixels belonging to AR and CH classes from March 1, 1997 until August 17, 2011. We observe a slight solar cycle dependence for the median intensity of the AR class. In the 171 Å channel, we observe a clear minimum around solar minimum, while the rest of the curve is rather flat. The 195 Å curve, on the other hand, features a slight maximum around the solar maximum and no notable trend in the rest of the curve. The median intensity of the AR class varies substantially more than that of the CH class.

Figure 6 presents the total intensity (in DN/s) of all pixels belonging to the AR and CH classes from March 1, 1997 until
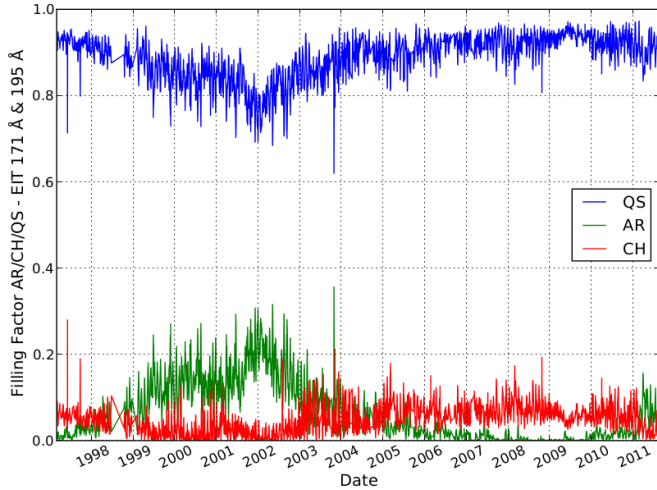
Fig. 7. Filling factor of AR, QS, and CH obtained by segmenting EIT 171 Å and EIT 195 Å images from March 1, 1997 until August 17, 2011.
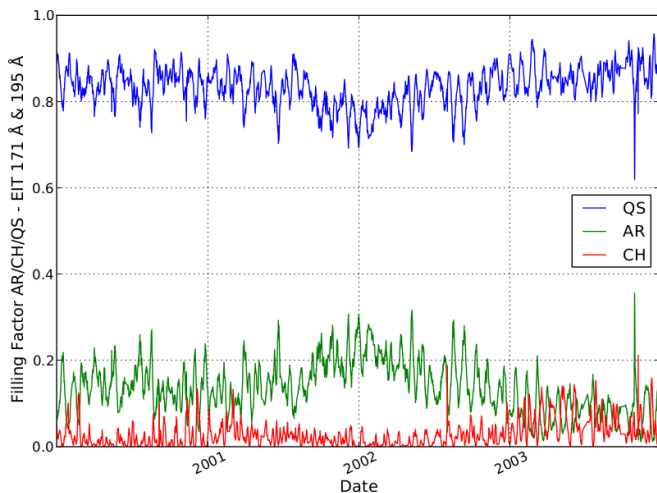

Fig. 8. Detail of the filling factor of AR, QS, and CH from January 1, 2000 until December 31, 2003.

August 17, 2011 (i.e., the sum of all those pixel values). The solar cycle is very obvious in the total intensity of AR with a clear double peak in 2000 and 2002 at 171 Å and 195 Å. The total intensity of CH was in antiphase with the solar cycle, as expected. The total intensity of AR displays a large variance.

Figure 7 presents the filling factor of all pixels belonging to AR, QS, and CH, respectively, from March 1, 1997 until August 17, 2011. The filling factor of the AR varies in clear correlation with the solar cycle, whereas the filling factors of both the QS and the CH vary in antiphase with the solar cycle, peaking at solar minimum.

Figure 8 zooms in on the filling factor of all pixels belonging to AR, QS, and CH in the period January 1, 2000−December 31, 2003. At these timescales, the solar rotation clearly shows up in AR, QS, and CH filling factors. It also becomes clear that the large variance in the AR intensity time series is mainly due to the variation in AR size with the solar rotation.

3.2. Segmentation of STEREO-EUVI images

Next, we turn to full Sun synoptic images from the Extreme Ultra Violet Imager EUVI (Wuelser et al. 2004) of the Sun Earth Connection Coronal and Heliospheric Investigation (SECCHI) imaging suite (Howard et al. 2008) onboard the STEREO spacecraft A(head) and B(ehind).

We have processed data from January 11, 2011 until March 31, 2011 at a cadence of one image per hour. Since the 171 Å dataset contained too many missing blocks, AR and CH classification was performed using only the 195 Å channel.

Similar to EIT, EUVI is also exposed to CRHs, and hence, the SPoCA2 algorithm also produces the best results. To improve the consistency of the results, the segmentation was done using a two-step approach. First, the SPoCA2 algorithm was run on the individual images. Second, the median values of the class centers and η_i parameters were computed, and an attribution was done on the individual images using these values.

Figure 9 shows an example of overlays of AR and CH maps onto the corresponding EUVI images. Every color corresponds to a different AR or CH as identified by the SPoCA-suite's region extraction algorithm. During the selected period, the STEREO A and B spacecraft were approximately 180° apart. Because of the dominant role of long-lived ARs and CHs returning to the same position after one solar rotation, the time series of filling factors should thus show roughly a 13.5 day lag between spacecraft A and B (corresponding to the observation of the same hemisphere on the Sun, observed with a time difference of 13.5 days). As can be seen from the plots in Fig. 10, this is indeed the case. In these plots, the results for STEREO B have been shifted 13.5 days ahead in time. Notice that some small periods of time were missing from the data for either of the spacecraft. For EUVI A, the period February 13−17 was missing, and for both EUVI A and B (before the time shift), the period February 27−March 2 was missing.

In the AR plot in Fig. 10a, the correlation is most apparent during the period from February 20 until March 31. The difference in magnitude of the filling factors for the two spacecraft is due to the relatively long amount of time between the corresponding observations and the high variability of AR filling factors.

In the CH plot in Fig. 10b, we can clearly see the solar rotation from the peaks around the 6th of February and March. These peaks were mostly due to a large CH in the northern hemisphere that stretched from high to low latitudes. Around the 6th of February, we can see that the filling factor, as observed by EUVI A, increased slightly from the filling factor, as observed by EUVI B 13.5 days earlier. The observations of the same set of CHs after one solar rotation yield similar filling factors for both spacecraft.

3.3. Segmentation of PROBA2-SWAP images

The Sun Watcher with Active Pixels and Image Processing SWAP (Berghmans et al. 2006; De Groof et al. 2008; Seaton et al. 2012) is a wide-field EUV solar imager onboard the PROBA2 spacecraft in low-Earth orbit. The PROBA2-SWAP observations consist of a series of images at 174 Å (Fe IX/X at $\log T \simeq 6.0$) taken with a cadence of roughly 100 s. The SWAP images are 1k × 1k with a linear pixel size of approximately 3.17 arcsec, meaning the instrument has a total field of view of approximately 54 × 54 arcmin^2.
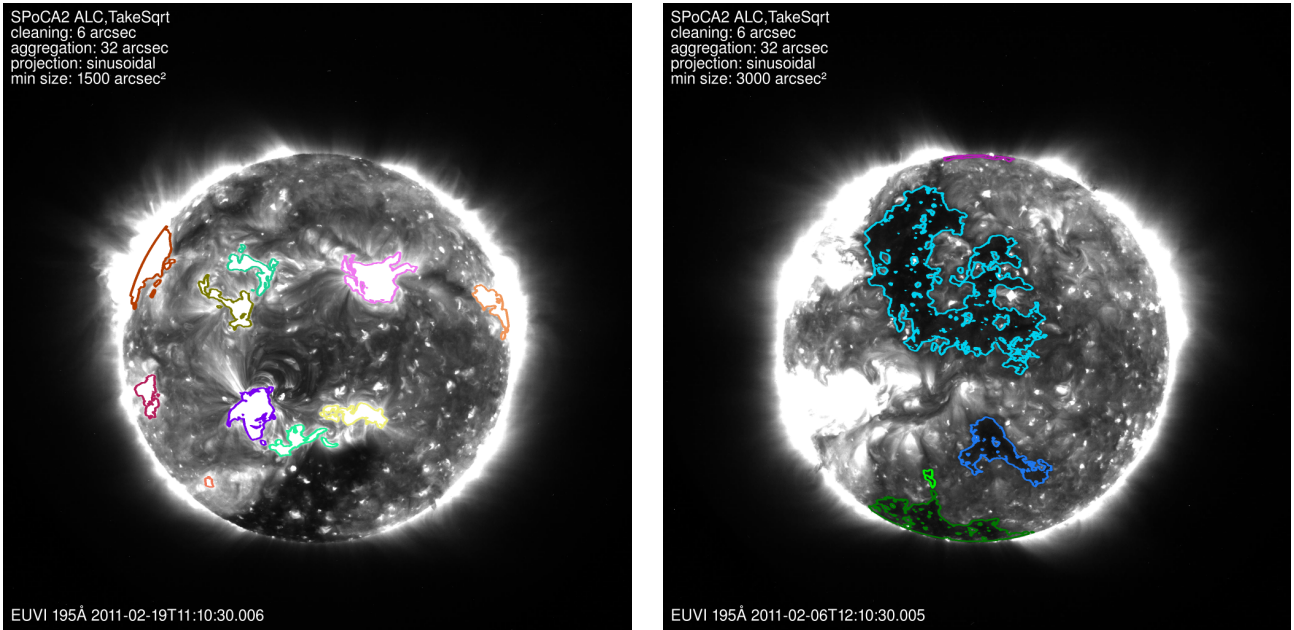
Fig. 9. Overlay of **a)** (*left*) an AR map onto the corresponding EUVI 195 Å image taken on February 19, 2011 around 11^h UT; and **b)** (*right*) a CH map onto the corresponding EUVI 195 Å image taken on February 6, 2011 around 12^h UT.
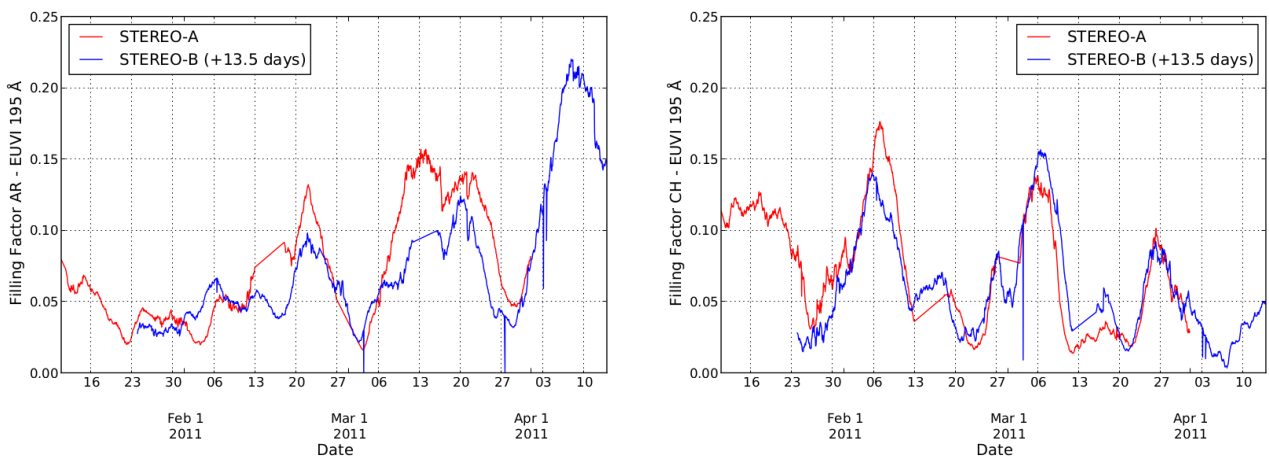

Fig. 10. Filling factor for **a)** (*left*) AR and **b)** (*right*) CH on EUVI A and B at 195 Å from January 11 until March 31, 2011.

From the SWAP data archive[10], we have selected one FITS file every six hours from October 1, 2010 until June 30, 2011. A SPoCA2 segmentation with six classes was run on this dataset to create a time series of the total intensity and raw area of ARs.

Figure 11 shows an example of the overlay of an AR map onto the corresponding SWAP image. Every color corresponds to a different AR as identified by the SPoCA-suite's region extraction algorithm. Note that the SPoCA-suite parameters for this analysis were selected in such a way that the detection of AR yields smaller regions than in our EIT analysis and corresponds to the AR core rather than the whole region.

Figure 12a shows the raw AR area and total AR intensity as detected by the SPoCA-suite on 174 Å SWAP images from October 1, 2010 to June 30, 2011. When comparing this to the corresponding time series of the daily International Sunspot Number (ISN) and the F10.7 radio flux in Fig. 12b, we observe a clear correlation, especially between the total AR intensity and

both ISN and F10.7. This suggests that the total intensity of ARs in narrow-band EUV imager channels is a good indicator of solar activity, which deserves to be studied in more detail.

3.4. Segmentation of SDO-AIA images

Since May 2010, the Atmospheric Imaging Assembly AIA (Lemen et al. 2012) on board SDO delivers 4k × 4k images of solar corona and continuum at a 10 s cadence and in 10 bandpasses. In this work, we consider the 171 and 195 Å bandpasses.

The SPoCA-suite has been running in near-real time on SDO-AIA data since September 2010 as part of the SDO Feature Finding Project. The resulting AR events are automatically ingested by the HEK, which provides catalogs of ARs and CHs containing properties such as location (through a bounding box or a chain code), area, and statistical moments of intensities. Contours can also be visualized through the isolsearch interface or within the JHelioviewer visualization software, as seen
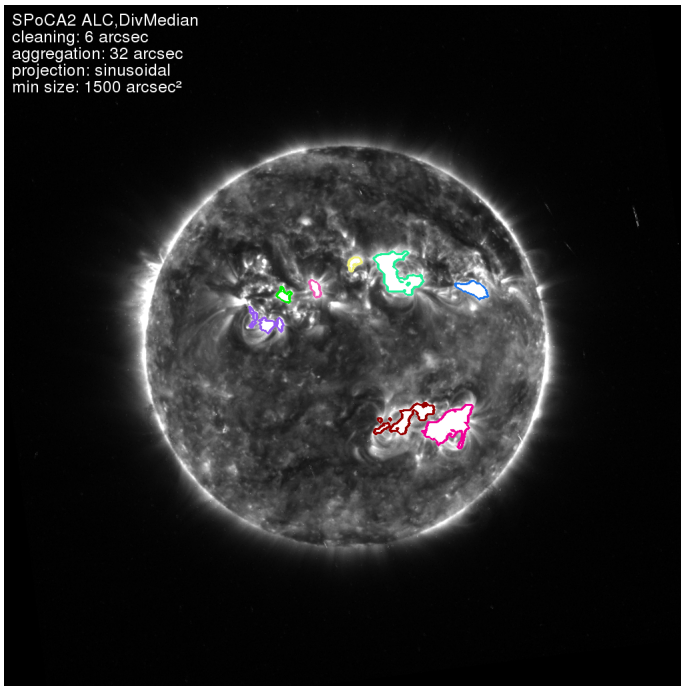
---

[10] http://proba2.sidc.be/data/SWAP

SPoCA2 ALC,DivMedian
cleaning: 6 arcsec
aggregation: 32 arcsec
projection: sinusoidal
min size: 1500 arcsec²

Fig. 11. Overlay of an AR map onto the corresponding SWAP 174 Å image taken on June 5, 2011 around 18[h] UT.

in Fig. 13 which shows a screenshot from the ESA JHelioviewer tool with AR and CH boundary overlay. A third way to access the images and their metadata is offered by the VSO.

To avoid prohibitive computation time on the 4k × 4k AIA images, iterations of the fuzzy clustering are done on the histogram of the preprocessed images rather than on the preprocessed images themselves. A bin size of 0.01 DN/s was found to be sufficiently small to produce accurate results. Histogram-based FCM yields the best results for extracting CHs out of the (almost noise free) 193 Å AIA images, whereas the histogram-based PCM2 algorithm is employed to detect ARs on AIA 171 Å and 193 Å images. The images were preprocessed by dividing by exposure time, correcting for limb brightness enhancement, and dividing by the median value.

Figure 14 shows an example of overlays of AR and CH maps onto the corresponding AIA images. Every color corresponds to a different AR or CH as identified by the SPoCA-suite's region extraction algorithm.

In Fig. 15, we present the AR, QS, and CH filling factor computed on a dataset of 12 months ranging from January 1 until December 31, 2011 at a twelve hour cadence. The solar rotation of about 27 days is noticeable (especially in the CH filling factor), but other periodicities (with a period of around 9 days) also seem to be present, similar to what was observed for CH by Temmer et al. (2007) based on GOES-SXI images. In this study, please note that we have selected SPoCA parameters such that the AR detections correspond to the AR cores, which results in much lower AR filling factors than in our EIT study.

4. Conclusions

We have introduced a set of segmentation procedures (called the SPoCA-suite) that allows users to retrieve active region (AR) and coronal hole (CH) properties from EUV images. The SPoCA-suite consists of the algorithms FCM, PCM, PCM2, SPoCA, and SPoCA2, and of histogram-based versions. The

main segmentation algorithms in the SPoCA-suite and their output were described in detail in Sect. 2, along with pre- and post-processing, region extracting, and tracking of individual regions over time.

The SPoCA-suite is the result of major improvements and extensions (see Sect. 2) on our previous work on the Spatial Possibilistic Clustering Algorithm (Barra et al. 2009). The main advances are robust solutions for two fundamental problems native to possibilistic algorithms (coincident clustering and the stability of η_i parameters), smooth variation in center values to enable automatic segmentation on a continuous stream of data, and more accurate segmentations near the solar limb due to a smoother limb brightness correction.

The SPoCA-suite provides a powerful tool for consistent automatic detection of AR, QS (quiet Sun), and CH. Moreover, it was demonstrated that the SPoCA-suite's automatic detection scheme enables the construction of large time series of properties pertaining to ARs or CHs, as observed by EUV imagers. We have presented and investigated such time series from four different EUV imagers (SOHO-EIT, STEREO-EUVI, PROBA2-SWAP, and SDO-AIA) in Sect. 3. These kinds of studies can be performed for other EUV imagers as well and opens the way for systematic large-scale surveys of AR and CH properties.

The SPoCA-suite allows users to identify and characterize both individual ARs or CHs and the total class of all AR or CH pixels in an EUV image. Individual regions can also be tracked over time. Masks, contours, contour overlays onto the original image, and overlay movies can be generated as output.

For any individual AR or CH, the SPoCA-suite provides the location of its barycenter and the corners of a bounding box. A number of characteristics (intensity moment statistics, area, and filling factor) are computed for every individual region and every total class.

The SPoCA-suite's automatic segmentation allowed us to construct long time series for EIT (1997-2011). The median 171 Å and 195 Å intensity in AR pixels (Fig. 5) shows a slight dependence in phase with the solar cycle, suggesting that ARs are slightly brighter in EUV in solar maximum than in solar minimum. In the 171 Å channel, we observe a clear minimum around the solar minimum, while the rest of the curve is flat. The 195 Å curve, on the other hand, features a slight maximum around solar maximum and no notable trend in the rest of the curve. The median intensity in AR pixels showed a much larger variance than the median intensity in CH pixels.

The total 171 Å and 195 Å intensity in AR pixels (Fig. 6) was clearly in phase with the solar cycle, even showing a double maximum in 2000 and 2002, whereas the total intensity of CH was in antiphase with the solar cycle. The filling factor of AR pixels (Fig. 7) shows a clear correlation with the solar cycle, whereas the filling factors of both QS and CH feature an anticorrelation with the solar cycle. Zooming in (2000−2003) on the filling factors of AR, QS, and CH (Fig. 8), we clearly observe the periodicity of the solar rotation. These filling factors can be included into (semi-)empirical models of the solar atmosphere, which can be employed to model the solar EUV irradiance (Haberreiter 2012).

For EUVI 195 Å, the SPoCA-suite was employed to study the filling factors for ARs and CHs from January until April 2011, when STEREO A and STEREO B were about 180 degrees apart. The signature of the solar rotation (due to long-lived ARs and CHs returning to the same position after one solar rotation) is obvious (Fig. 10). Shifting EUVI B's AR and CH filling factors by 13.5 days (about half a solar rotation)
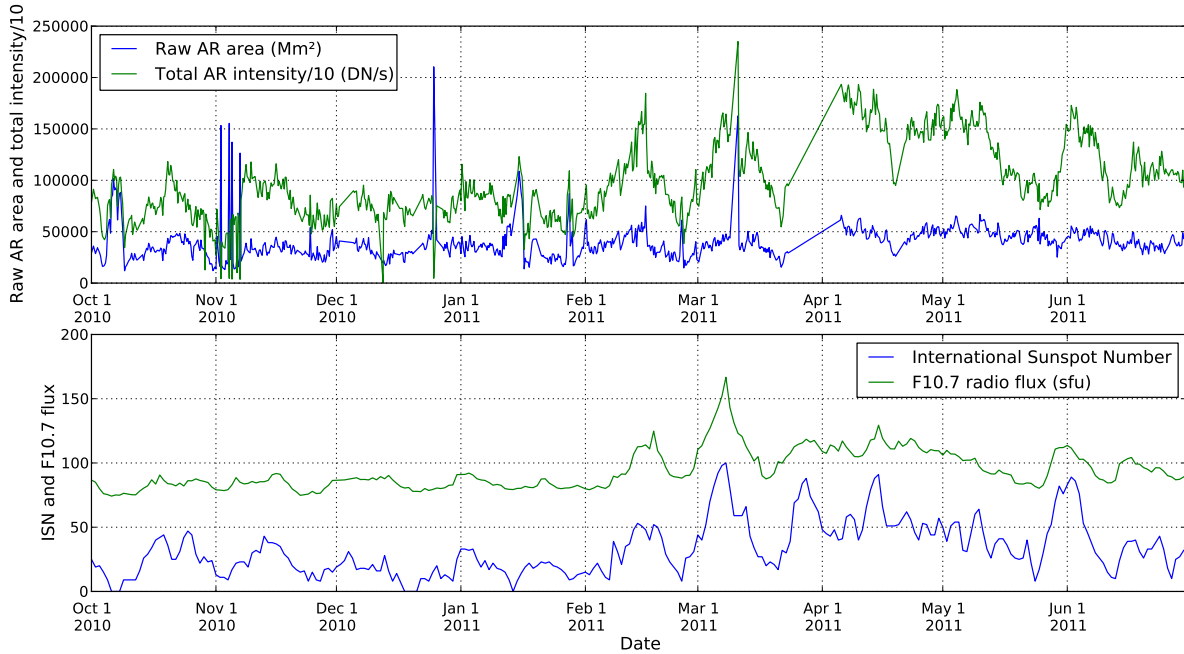
Fig. 12. a) (*top*) Raw AR area and total AR intensity from 174 Å SWAP images; and **b)** (*bottom*) Daily International Sunspot Number and F10.7 radio flux from October 1, 2010 to June 30, 2011.
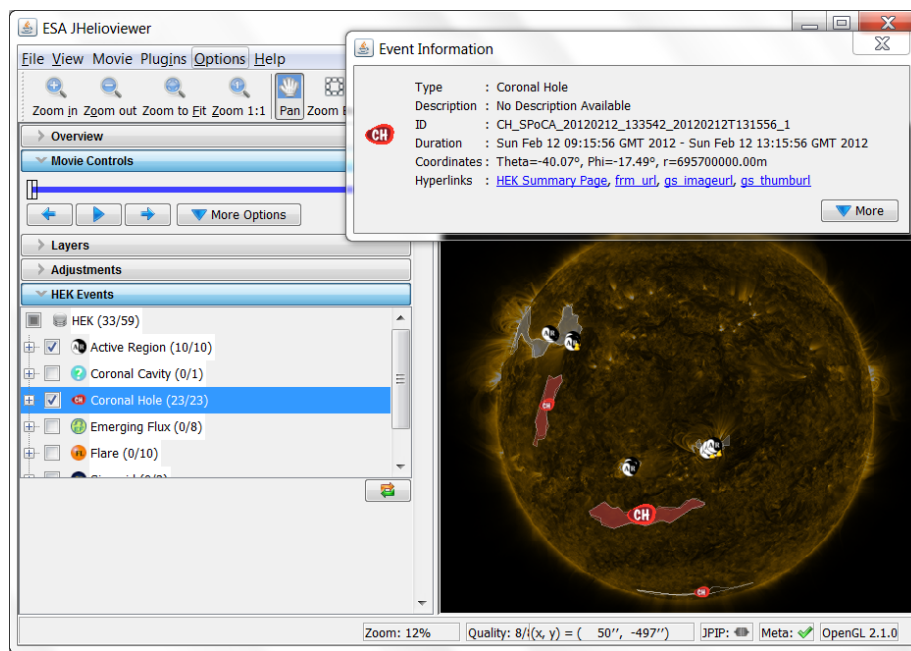

Fig. 13. Screenshot from the ESA JHelioviewer tool. The picture on the right displays the AIA 171 Å image taken on February 12, 2012 at 9:15:56 UT with AR and CH location and chain code information that are recorded in the HEK. An event information window pops up when clicking on an event or feature (here the large CH located in the Southern hemisphere).

results in filling factors very similar to those observed by EUVI A, which was expected since EUVI A and the shifted EUVI B correspond to the same hemisphere on the Sun, observed with a time difference of 13.5 days.

For SWAP, the total 174 Å AR intensity and the total AR area (Fig. 12) were extracted over the period October 2010–June 2011. The resulting time series are clearly correlated with both the daily International Sunspot Number and the F10.7 radio flux, and hence provide a novel and useful proxy for solar activity.

This kind of time series will soon be available in the Solar Timelines viewer for AFFects (STAFF), a dedicated viewer for solar activity, solar wind, and geomagnetic timelines developed at the Royal Observatory of Belgium within the European FP7 project AFFECTS.

The SPoCA-suite was also applied to AIA 171 and 193 Å, and the evolution of AR, QS, and CH filling factors was analyzed for the period January–December 2011 on a twelve hour cadence (Fig. 15). The effect of the solar rotation is again conspicuous in
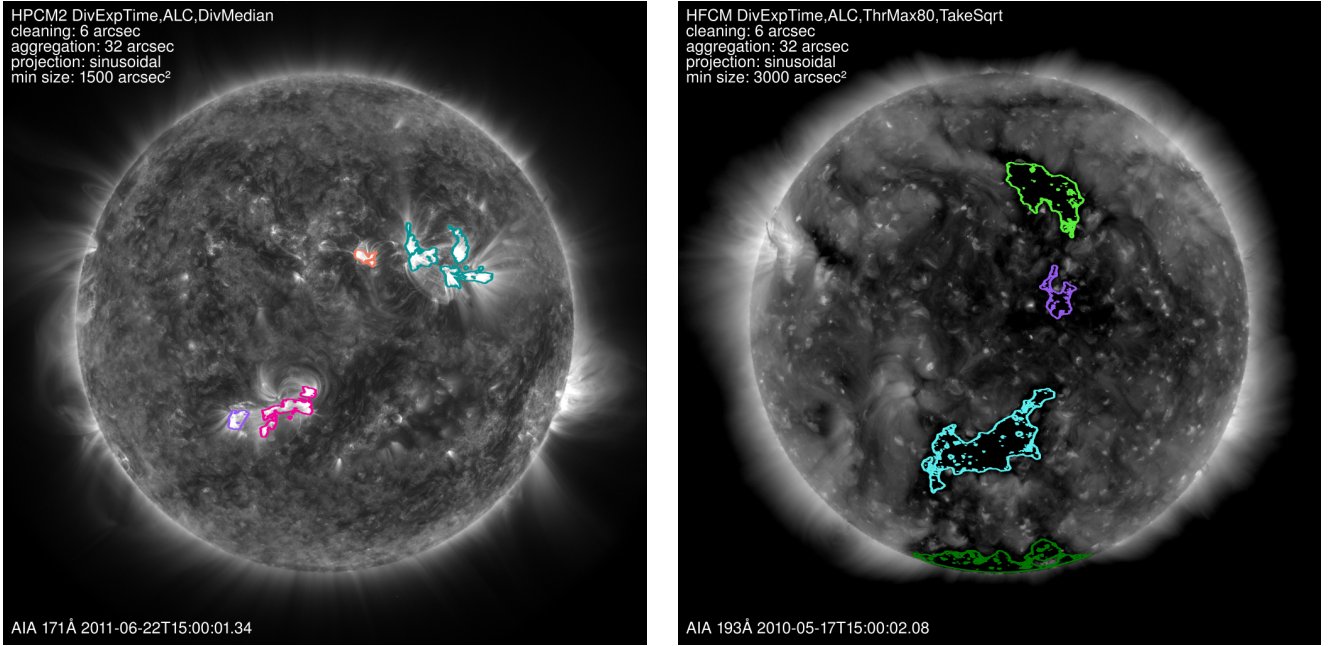
Fig. 14. a) (*left*) Overlay of an AR map created using AIA 171 and 193 Å onto the corresponding AIA 171 Å image taken on June 22, 2011 around 15^h UT. **b)** (*right*) Overlay of a CH map created using AIA 193 Å onto the corresponding AIA 193 Å image taken on May 17, 2010 around 15^h UT.
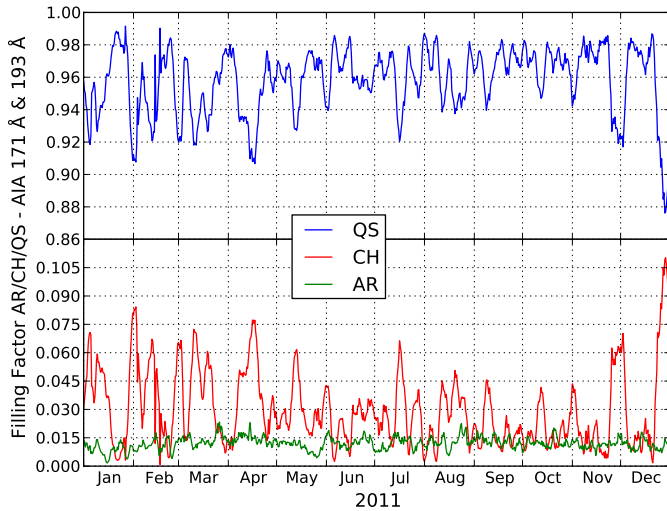

Fig. 15. Filling factor of AR, QS, and CH obtained from segmenting AIA 171 and 193 Å images from January 1 until December 31, 2011.

these data, but a period of around 9 days also shows up, which is similar to what was observed for CH by Temmer et al. (2007) based on GOES-SXI images.

One of the SPoCA-suite algorithms has also been implemented in the modules of the HEK. It provides catalogs of ARs and CHs containing properties, such as localization (through a bounding box or a chain code), area, and moments of intensities, based on AIA images. Contours can also be visualized through the isolsearch interface, within the JHelioviewer visualization software, or through the VSO.

Avenues for future research include improving criteria for distinguishing between filaments and CHs, segmenting in 3D (the third dimension being time) to improve the accuracy of

tracking, and, finally, tailoring the segmentation to the needs of EUV irradiance reconstruction models.

Appendix A: Constraints on the regularization parameter η_i

All algorithms based on Possibilistic C-means strongly depend on the choice of the regularization parameter η_i. A various number of elaborated procedures have been proposed in the literature, as seen in Krishnapuram & Keller (1996). An intuitive choice is to compute η_i as the intraclass dispersion, as in Eq. (7). Problems arise, however, when the underlying classes have a widely different intraclass variance. For example, the highly variable AR class on EUV images may include the darkest part of what should be the CH class in the final segmentation.

To understand this phenomenon, consider two classes c_1 and c_2 with centers b_1 and b_2. Suppose $b_{1q} < b_{2q}, \forall q = 1, \ldots, p$, where p is the dimension of the feature vectors. Under certain circumstances, we show that u_{2j} can exceed u_{1j} for values of x_j where $x_{jq} < b_{1q}, \forall q = 1, \ldots, p$.

Let us first determine the locus of points x_j, where $u_{1j} = u_{2j}$. From Eq. (8) we find that $u_{1j} = u_{2j}$ if and only if

$$\frac{d^2(x_j, b_1)}{\eta_1} = \frac{d^2(x_j, b_2)}{\eta_2}. \tag{A.1}$$

If $\eta_1 = \eta_2$, the solution is a hyperplane through the middle of b_1 and b_2, and there are no undesired effects. In the case of $\eta_1 \neq \eta_2$ and for the Euclidean distance, the above is the equivalent to saying that x_j lies on a circle with center

$$c = \frac{\eta_2 b_1 - \eta_1 b_2}{\eta_2 - \eta_1} \tag{A.2}$$

and radius

$$r = \frac{\sqrt{\eta_1 \eta_2}}{|\eta_2 - \eta_1|} d(b_1, b_2). \tag{A.3}$$

We consider only the case $\eta_1 < \eta_2$, as this is the case for the classes CH, QS, and AR when they arise in EUV images. In this case, c lies relatively close to b_1 at the opposite side from b_2. All points x_j inside the circle satisfy $u_{1j} > u_{2j}$ and hence, belong to class 1. All points x_j outside the circle satisfy $u_{2j} > u_{1j}$, and hence, are classified as belonging to class 2 instead of class 1. This is unwanted behavior for those points for which some $x_{jq} < b_{1q}$. To avoid this situation, we can select η_i-values in such a way that the circle center c lies below all coordinate axes. The result is that for all points x_j in the circle, all positive points below it are also in the circle. Hence, whenever a point x_j belongs to class 1, all points below x_j also belong to class 1. So we want $c_q < 0, \forall q = 1, \ldots, p$, which is equivalent to the following conditions on η_1 and η_2:

$$\frac{\eta_2}{\eta_1} < \frac{b_{2q}}{b_{1q}}, \forall q = 1, \ldots, p. \tag{A.4}$$

The same considerations apply to the situation where $\eta_1 < \eta_2 < \eta_3$ etc.

Appendix B: Computation of the CH chain code

Chain coding aims at representing the boundary of an object in digitized images. It is based on the idea of following the outer edge of the object and storing the direction when traveling along the boundary (Freeman 1961). In the SPoCA-suite, we use a representation of the chain code with eight directions, as is commonly done in the literature, see Castleman (1996). This has the same length as the perimeter of the object under consideration, which in many cases is too long.

In a second step, we thus find a polygonal approximation to the perimeter that has a limited number of edges and for which the distance from any point in the perimeter to the polygon does not exceed a given accuracy.

We use the algorithm described in Douglas & Peucker (1973), which proceeds as a recursive refinement. The main axis of the contour is first extracted, providing the first two vertices. Each polygon edge is then recursively split by introducing a new vertex at the most distant associated contour point until the desired accuracy is reached. More precisely, the algorithm runs as follows:

1. Initialize the polygon with points p_1 and p_2 of the perimeter that are farthest away from each other.
2. Let $i = 3$.
3. For each segment in the polygon, find the point on the perimeter between the points that have the farthest distance to the polygonal linesegment. If this distance is larger than a threshold, mark the point with a label p_i.
4. Renumber the points, so that they are consecutive.
5. Increment i by one.
6. If no points have been added, then break, otherwise go to 3.

Within the HEK, a maximal number of 100 points is considered to be sufficient.

References

Abramenko, V., Yurchyshyn, V., & Watanabe, H. 2009, Sol. Phys., 260, 43
Anscombe, F. J. 1948, Biometrika, 35, 246
Aschwanden, M. J. 2010, Sol. Phys., 262, 235
Baker, J., Campbell, D., & Bodnarova, A. 2003, in Proc. ASC'2003, Artificial Intelligence and Soft Computing, 385
Barra, V., Delouille, V., & Hochedez, J. 2008, Adv. Space Res., 42, 917
Barra, V., Delouille, V., Kretzschmar, M., & Hochedez, J.-F. 2009, A&A, 505, 361
Benkhalil, A., Zharkova, V., Ipson, S., & Zharkov, S. 2003, in Proc. of the AISB'03 Symposium on Biologically-inspired Machine Vision, Theory and Application, University of Wales, Aberystwyth, 7th−11th April, 66
Benkhalil, A., Zharkova, V. V., Zharkov, S., & Ipson, S. 2006, Sol. Phys., 235, 87
Berghmans, D., Hochedez, J., Defise, J., et al. 2006, Adv. Space Res., 38, 1807
Bernasconi, P. N., Rust, D. M., & Hakim, D. 2005, Sol. Phys., 228, 97
Bezdek, J. 1981, Pattern recognition with fuzzy objective function algorithms (New-York: Plenum Press)
Bezdek, J., Hall, L. O., Clark, M., Goldof, D., & Clarke, L. 1997, Stat. Meth. Med. Res., 6, 191
Castleman, K. 1996, Digital Image Processing (New Jersey: Prentice Hall)
Colak, T., & Qahwaji, R. 2008, Sol. Phys., 248, 277
Colak, T., & Qahwaji, R. 2009, Space Weather, 7, 6001
Colak, T., & Qahwaji, R. 2013, Sol. Phys., 283, 143
Curto, J. J., Blanca, M., & Martínez, E. 2008, Sol. Phys., 250, 411
De Groof, A., Berghmans, D., Nicula, B., et al. 2008, Sol. Phys., 249, 147
de Toma, G. 2011, Sol. Phys., 274, 195
de Toma, G., & Arge, C. N. 2005, in Large-scale Structures and their Role in Solar Activity, eds. K. Sankarasubramanian, M. Penn, & A. Pevtsov, ASP Conf. Ser., 346, 251
Delaboudinière, J., Artzner, G. E., Brunaud, J., et al. 1995, Sol. Phys., 162, 291
Douglas, D., & Peucker, T. 1973, Journal Cartographica: The International Journal for Geographic Information and Geovisualization, 10, 112
Dudok de Wit, T. 2006, Sol. Phys., 239, 519
Freeman, H. 1961, Proc. IRE Translation Electron Computer, 260 (New York)
Gan, G., Ma, C., & Wu, J. 2007, Data clustering: theory, algorithms, and applications, Soc. Industr. Appl. Math., 20
Gurman, J. B., Hill, F., Suàrez-Solà, F., et al. 2012, in AAS Meeting Abstracts, 220, 201.24
Haberreiter, M. 2012, in IAU Symp. 286, eds. C. H. Mandrini & D. F. Webb, 97
Henney, C. J., & Harvey, J. W. 2005, in Large-scale Structures and their Role in Solar Activity, eds. K. Sankarasubramanian, M. Penn, & A. Pevtsov, ASP Conf. Ser., 346, 261
Higgins, P. A., Gallagher, P. T., McAteer, R. T. J., & Bloomfield, D. S. 2011, Adv. Space Res., 47, 2105
Hill, F., Bogart, R., Davey, A., et al. 2004, in Optimizing Scientific Return for Astronomy through Information Technologies, Proc. SPIE, 5493, 163
Howard, R. A., Moses, J. D., Vourlidas, A., et al. 2008, Space Sci. Rev., 136, 67
Hurlburt, N., Cheung, M., Schrijver, C., et al. 2012, Sol. Phys., 275, 67
Jones, H. 2004, in Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference, KES 2004, eds. M. G. Negoita, R.J. Howlett, & L. C. Jain, Computer Science, 3215, 433
Kestener, P., Conlon, P. A., Khalil, A., et al. 2010, ApJ, 717, 995
Kirk, M. S., Pesnell, W. D., Young, C. A., & Hess Webber, S. A. 2009, Sol. Phys., 257, 99
Krieger, A. S., Timothy, A. F., & Roelof, E. C. 1973, Sol. Phys., 29, 505
Krishnapuram, R., & Keller, J. 1993, IEEE Trans. Fuzzy Systems, 1, 98
Krishnapuram, R., & Keller, J. 1996, IEEE Trans. Fuzzy Systems, 4, 385
Krista, L. D., & Gallagher, P. T. 2009, Sol. Phys., 256, 87
Lefebvre, S., & Rozelot, J. 2004, Sol. Phys., 219, 25
Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, Sol. Phys., 275, 17
Malanushenko, O. V., & Jones, H. P. 2005, Sol. Phys., 226, 3
Martens, P. C. H., Attrill, G. D. R., Davey, A. R., et al. 2012, Sol. Phys., 275, 79

McAteer, R. T. J., Gallagher, P. T., Ireland, J., & Young, C. A. 2005, Sol. Phys., 228, 55

Melgani, F., Hashemy, B. A., & Taha, S. 2000, IEEE Trans. Geoscience and Remote Sensing, 38, 287

Müller, D., Fleck, B., Dimitoglou, G., et al. 2009, Comput. Sci. Eng., 11, 38

Nguyen, S. H., Nguyen, T. T., & Nguyen, H. S. 2005, in Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, eds. D. Slezak, J. Yao, J. F. Peters, W. Ziarko, & X. Hu (Berlin/Heidelberg: Springer), Lect. Notes Comput. Sci., 3642, 263

Nieniewski, M. 2002, in Proc. SOHO 11 Symposium on From Solar Min to Max: Half a Solar Cycle with SOHO, eds. N. E. P. D. A. Wilson, Noordwijk, 323

Obridko, V. N., Shelting, B. D., Livshits, I. M., & Asgarov, A. B. 2009, Sol. Phys., 260, 191

Pérez-Suárez, D., Higgins, P. A., McAteer, R. T. J., Bloomfield, D. S., & Gallagher, P. T. 2011, in Applied Signal and Image Processing: Multidisciplinary Advancements, eds. R. Qahwaji, R. Green, & E. Hines (Hershey, Pennsylvania: IGI Global), 207

Pettauer, T., & Brandt, P. 1997, Sol. Phys., 175, 197

Philipps, W., Velthuizen, R., & S.Phuphanich. 1995, Magn. Reson. Imaging, 13, 277

Preminger, D., Walton, S., & Chapman, G. 1997, Sol. Phys., 171, 303

Qahwaji, R., & Colak, T. 2006, I. J. Comput. Appl., 13, 9

Rangsanseri, Y., & Thitimajshima, P. 1998, in Proc. IGARSS, 5, 2507

Revathy, K., Lekshmi, S., & Nayar, S. R. P. 2005, Sol. Phys., 228, 43

Saar, S. H., & Farid, S. 2011, in AAS/Solar Physics Division Abstracts #42, 2121

Scholl, I. F., & Habbal, S. R. 2008, Sol. Phys., 248, 425

Seaton, D. B., Berghmans, D., Nicula, B., et al. 2012, Sol. Phys., 217

Snyder, J. P. 1987, Map Projections: A Working Manual, USGS Numbered Series No. 1365 (Geological Survey, US)

Steinegger, M., Bonet, J., & Vazquez, M. 1997, Sol. Phys., 171, 303

Steinegger, M., Bonet, J., Vazquez, M., & Jimenez, A. 1998, Sol. Phys., 177, 279

Temmer, M., Vršnak, B., & Veronig, A. M. 2007, Sol. Phys., 241, 371

Trauwaert, E., Kaufman, L., & Rousseeuw, P. 1991, Fuzzy Sets and Systems, 42, 213

Turmon, M., Pap, J. M., & Mukhtar, S. 2002, ApJ, 568, 396

Verbanac, G., Vršnak, B., Veronig, A., & Temmer, M. 2011, A&A, 526, A20

Verbeeck, C., Higgins, P., Colak, T., et al. 2011, Sol. Phys., 283, 67

Vršnak, B., Temmer, M., & Veronig, A. M. 2007, Sol. Phys., 240, 315

Watson, F., Fletcher, L., Dalla, S., & Marshall, S. 2009, Sol. Phys., 260, 5

Worden, J., Woods, T., Neupert, W., & Delaboudiniere, J. 1999, ApJ, 965

Wuelser, J.-P., Lemen, J. R., Tarbell, T. D., et al. 2004, in SPIE Conf. Ser. 5171, eds. S. Fineschi, & M. A. Gummin, 111

Zharkov, S., Zharkova, V. V., & Ipson, S. S. 2005, Sol. Phys., 228, 377