The Astrophysical Journal, 930:118 (11pp), 2022 May 10 https://doi.org/10.3847/1538-4357/ac5f43

Identification of Coronal Holes on AIA/SDO Images Using Unsupervised Machine Learning

Fadil Inceoglu[1,2,3], Yuri Y. Shprits[1,4,5], Stephan G. Heinemann[6], and Stefano Bianco[1]
[1] GFZ German Research Centre for Geosciences, Potsdam, Germany; fadil@gfz-potsdam.de
[2] Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA
[3] National Centers for Environmental Information, National Oceanographic and Atmospheric Administration, Boulder, CO, USA
[4] Institute for Physics and Astronomy, University of Potsdam, Potsdam, Germany
[5] Department of Earth, Planetary, and Space Science, University of California, Los Angeles, CA, USA
[6] Max Planck Institute for Solar System Research, Goettingen, Germany
Received 2022 January 28; revised 2022 March 10; accepted 2022 March 19; published 2022 May 10

Abstract

Through its magnetic activity, the Sun governs the conditions in Earth's vicinity, creating space weather events, which have drastic effects on our space- and ground-based technology. One of the most important solar magnetic features creating the space weather is the solar wind that originates from the coronal holes (CHs). The identification of the CHs on the Sun as one of the source regions of the solar wind is therefore crucial to achieve predictive capabilities. In this study, we used an unsupervised machine-learning method, *k*-means, to pixel-wise cluster the passband images of the Sun taken by the Atmospheric Imaging Assembly on the Solar Dynamics Observatory in 171, 193, and 211 Å in different combinations. Our results show that the pixel-wise *k*-means clustering together with systematic pre- and postprocessing steps provides compatible results with those from complex methods, such as convolutional neural networks. More importantly, our study shows that there is a need for a CH database where a consensus about the CH boundaries is reached by observers independently. This database then can be used as the "ground truth," when using a supervised method or just to evaluate the goodness of the models.

Unified Astronomy Thesaurus concepts: Solar coronal holes (678); Detection (1911); Space weather (2037)

1. Introduction

The Sun is a magnetically active star that shows various magnetic activity structures extending from its surface to its higher atmospheric layers, such as bipolar active regions (ARs) on the photosphere, filaments in the chromosphere, and coronal holes (CHs) in its corona. Through its magnetic activity, the Sun governs the conditions in the vicinity of Earth and throughout the heliosphere, which creates space weather and space climate. Space weather is defined as the effects of the solar wind and solar eruptive phenomena, such as flares and coronal mass ejections (CMEs), on Earth's magnetosphere, ionosphere, and thermosphere (Schwenn 2006). The space weather conditions have drastic effects our space- and ground-based technology (Eastwood et al. 2017).

One of the most important solar magnetic features creating the space weather and in turn affecting the Earth is the solar wind. The observations revealed that there are three different types of solar wind: (i) steady fast solar winds originate in the CHs, (ii) unsteady slow winds from opening magnetic loops and active regions, and (iii) transient winds from CMEs (Marsch 2006). The identification of the CHs on the Sun as one of the source regions of the solar wind (Wilcox 1968) that creates space weather and in turn influences our space- and ground-based technology is therefore crucial to achieve predictive capabilities.

As the source regions of the steady fast solar winds, CHs are identified as regions of low-density collisionless plasma that is generally located above inactive parts of the Sun, where open magnetic field lines extend throughout the heliosphere (Schwenn 2006; Cranmer 2009). The magnetic field inside a CH is known to be more unipolar and the CHs show sharp and/or diffuse transition on the boundaries between them and their surroundings (Cranmer 2009). The temporal evolution of the CH as well as the area they cover on the Sun depends on the solar activity cycle, also known as the Schwabe cycle (Schwabe 1844). During the minimum phase of a solar cycle, the CHs are observed to be larger and located mainly on the solar polar caps. On the inclining phase of a cycle, the CHs are observed to be present at any latitude and to be short lived. During solar maximum, the CHs are smaller and only exist around midlatitudes, while on the declining phase of the solar cycle there are more long-lived CHs at lower latitudes and they form closer to the solar equator as the cycle progresses (Hewins et al. 2020). Additionally, during the inclining and declining phases of a solar cycle, the CHs can evolve into structures extending from a solar pole to the solar equator.

As CHs have lower densities and temperatures, and hence the lowest emission in UV and X-ray in comparison to their surrounding environment consisting of active regions and quiet Sun, they appear as dark regions in solar images in wavelengths around 194 Å whether they are on-disk or off-limb CHs (Cranmer 2009).

Detection of CHs is done by eye on the He I 10830 Å near-infrared absorption line triplet (Harvey & Recely 2002), using histogram-based intensity thresholding on 193 and 195 Å passband images of the Sun from the Atmospheric Imaging Assembly (AIA; Lemen et al. 2012) on the Solar Dynamics Observatory (SDO; Pesnell et al. 2012) and the Extreme

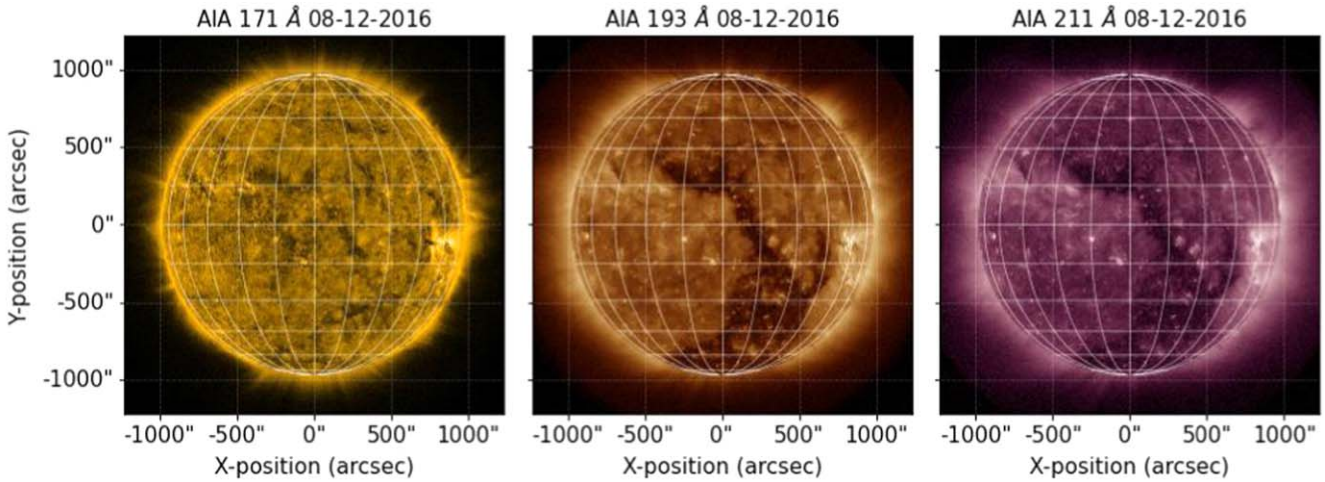

Figure 1. Passband images of the Sun in 171 Å (the left panel), 193 Å (the middle panel), and 211 Å (the right panel) taken by the AIA/SDO on 2016 December 8 at 00:00 UT.

ultraviolet Imaging Telescope (EIT; Delaboudinière et al. 1995) on the Solar and Heliospheric Observatory (SOHO), respectively (CHARM; Krista & Gallagher 2009). Additionally, an automated method for detection and segmentation of CHs based on multithermal intensity segmentation using 171, 193, and 211 Å passband images of the Sun from the AIA/SDO (CHIMERA; Garton et al. 2018), and a semiautomated method based on intensity threshold that is modulated by the intensity gradient of a CH have been developed (CATCH; Heinemann et al. 2019).

There are also methods based on supervised and unsupervised machine-learning (ML) methods. Verbeeck et al. (2014) developed a set of segmentation procedures based on the spatial possibilistic clustering algorithm (SPoCA) to detect CHs in an unsupervised ML fashion. Identified ARs and CHs by this algorithm are uploaded to the event catalogs in the Heliophysics Event Knowledge (HEK) database (Hurlburt et al. 2012). Illarionov & Tlatov (2018) used convolutional neural networks (CNNs; Schmidhuber 2014; Lecun et al. 2015) based on the U-Net architecture (Ronneberger et al. 2015) to identify CHs on solar images at 193 Å passband images of the Sun from AIA/SDO. They trained their network using binary maps from the Kislovodsk Mountain Astronomical Station. Recently, Jarolim et al. (2021) utilized CNNs based on a progressively growing architecture using data from all seven channels of AIA/SDO (94, 131, 171, 193, 211, 304, and 335 Å) as well as line-of-sight magnetograms from the Helioseismic and Magnetic Imager (HMI; Scherrer et al. 2012) on the SDO. For their network, the authors used binary maps from manually reviewed SPoCA-CH data (Delouille et al. 2018).

In this study, we utilize a pixel-wise *k*-means algorithm, which is an unsupervised ML method, to detect CHs based on 171, 193, and 211 Å passband images from the AIA/SDO. To achieve this objective, we used data from each channel in different combinations, and compared results from each combination to each other as well as to those from CATCH and the HEK data to calculate their performances. We first describe the data used in this study in Section 2 and explain the analyses and present our results in Section 3. We discuss the results and conclude in Section 4.

2. Data

To detect the CHs on the solar corona, we use passband data with 2 s exposure from AIA/SDO in wavelengths 171, 193, and 211 Å in different combinations (Figure 1). The AIA telescope on the SDO takes passband measurements of the Sun every 12 s in full disk with a spatial resolution of 4096×4096 pixels, and each pixel corresponds to $0\farcs6$ on the solar disk leading to a spatial resolution of $1\farcs5$ (Lemen et al. 2012). These three EUV bandpasses are centered on specific spectral emission lines of Fe IX for 171 Å, Fe XII, XXIV for 193 Å, and Fe XIV for 211 Å, which covers the temperature range from 6×10^5 to 2×10^6 K, corresponding to the upper transition region, quiet corona (171 Å), corona and hot flare plasma (193 Å), and active-region corona (211 Å) (Lemen et al. 2012).

3. Analyses and Results

3.1. Preprocessing Data

To detect the CHs, we use solar images taken by AIA/SDO in passband images in wavelengths 171, 193, and 211 Å in different configurations. We also study the most efficient wavelength or configuration of wavelengths to identify the CHs. To achieve this, we compare our CH binary maps with those from the CATCH. We also compared the CH polygons provided by the HEK with the CATCH binary maps to have a baseline with which we compare our results. The CATCH binary maps are selected from the last two months of each year in a time range from 2010 November to 2016 December, extending through solar cycle 24. The CATCH data in this period is reliable with minimal uncertainties. The total 237 CATCH CH binary maps consist of only contributions from the longitudinal range of $[-400, 400]$ arcsec in helioprojective coordinates as in this region the CHs can be identified more robustly (Jarolim et al. 2021). We also imported CH polygons from the HEK database for the same dates as the CATCH maps, and converted them into binary maps.

In total, we analyze 237 days of data. For each date, we import the Level 1 data in 171, 193, and 211 Å wavelengths and preprocess them using *aiapy* (Barnes et al. 2020a, 2020b) and *SunPy* (The SunPy Community et al. 2020; Mumford et al. 2021) python packages. This step consists of correcting the data for instrument degradation, for pointing and observer

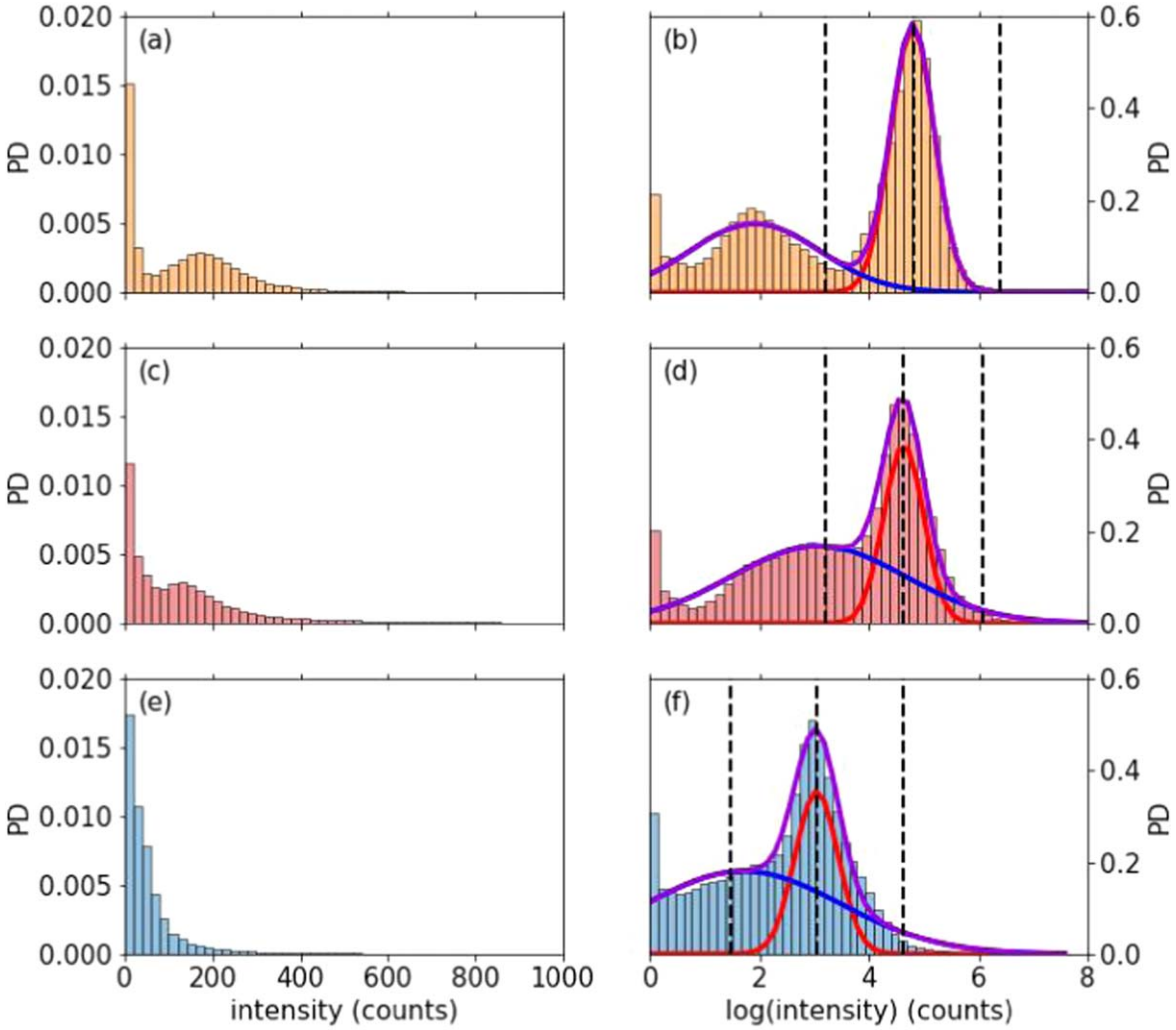

Figure 2. Probability densities of AIA/SDO 171 Å (top panel), 193 Å (middle panel), and 211 Å (bottom panel) intensities of the solar disk on 2016 December 8 at 00:00 UT. The left panels show the probability densities of the preprocessed data, while the right panels show probability densities of the postprocessed data. The vertical dashed lines show the mean (μ) and $\mu \pm 4\sigma$ values calculated to identify the threshold values.

location. Following these corrections, we register and align the data and normalize it so it has units of counts/pixels/seconds. Following these corrections, we correct the passband images for limb brightening using an annulus limb brightening correction approach (Verbeeck et al. 2014). We then deconvolve the passband images using the instrument point-spread function for each wavelength, and rescaled them to 1024×1024 using the spline method. As the final step, we lognormal transformed the data.

Following these steps, we created histograms of each data set to determine the lower- and upper-threshold values. Determining these values allows us to increase the contrast in the data. To avoid using any arbitrary values for these thresholds and to have a more systematic approach for determining these values, we fit a bimodal Gaussian curve to each histogram (Figure 2), where it is possible. For some dates, however, it was not possible to fit a bimodal Gaussian fit. For these dates, we used a unimodal Gaussian fit. Using the obtained parameters of the Gaussian fits, we calculated the lower- and upper-threshold values based on the mean and standard deviation values of the higher peak (the right panels of Figure 2), because the lower peak represents the CH pixels (Heinemann et al. 2019). For each date in the data set, we calculate a lower-threshold value for each wavelength based on $(\mu - 4\sigma)$, while the upper-threshold value is determined based on $(\mu + 4\sigma)$. Values below (above) the lower-threshold (upper-threshold) value are stacked to have only one value that is the threshold value.

We then investigate the temporal variations in the calculated mean (μ) and the lower-threshold values ($\mu - 4\sigma$) (Figure 3). The μ values of 193 and 211 Å passband images show variations in phase with the solar cycle, while the μ values of 171 Å do not show such a trend (Figure 3(a)). The μ values for each passband image also show day-to-day fluctuations. Similarly, the lower-threshold values show day-to-day fluctuations as well. These fluctuations have a wider range for the threshold values calculated for the 211 Å passband images especially during the maximum phase of the solar cycle, while the other two channels do not exhibit such wide fluctuations (Figure 3(b)). An important feature to note is the "negative" threshold values found for the 211 Å passband images. There are 27 days where the lower thresholds are negative values. However, as this does not have a physical meaning, the threshold values for these days were accepted as zero. The reason for the negative values come from the underlying shape of the Gaussians.

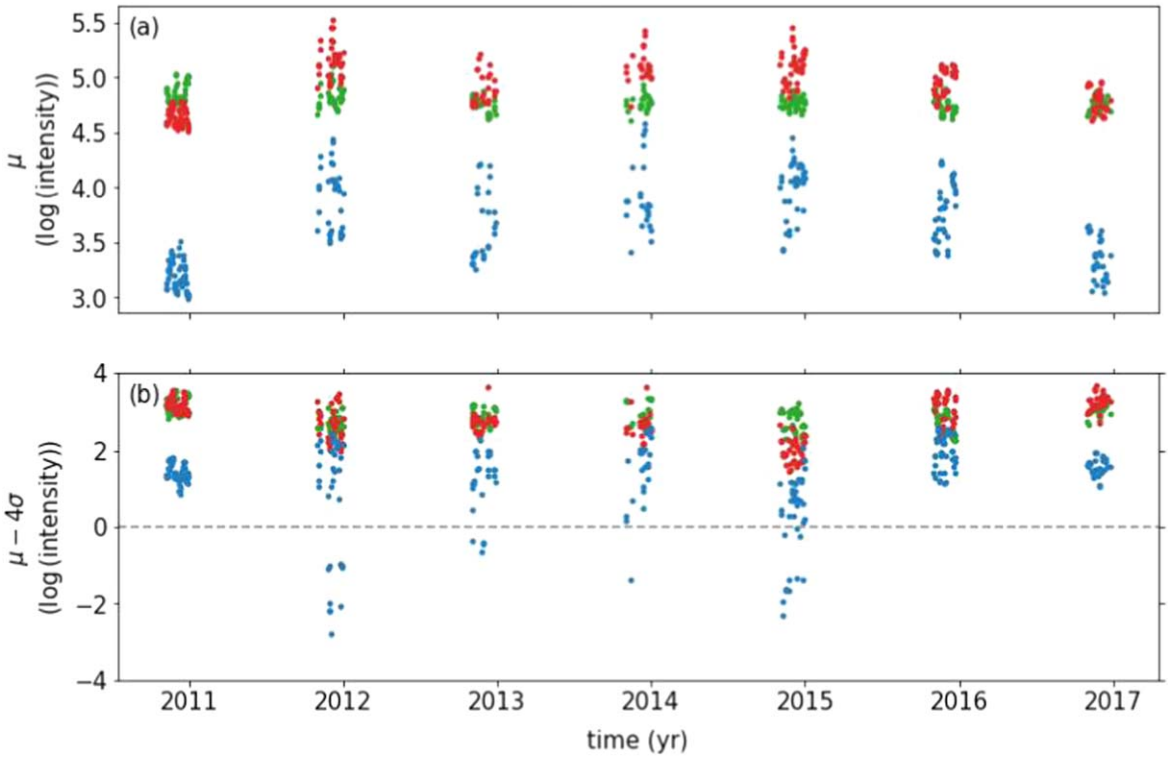

Figure 3. Calculated mean (μ) (a) and lower-threshold values ($\mu - 4\sigma$) (b) for AIA/SDO 171 Å (green), 193 Å (red), and 211 Å (blue) passband images for the study period. Note that there are 27 points below zero, meaning that no lower-threshold value could be calculated; therefore, no thresholding was applied to the 211 Å passband data on these dates.

3.2. *Pixel-wise Clustering the Images Using the* k*-means Algorithm*

After increasing the contrast in each image based on their individual mean and standard deviation values, we created four different data sets: (i) 193 Å image, (ii) 211 Å image, (iii) 193 and 211 Å composite image (two-channel composite (2CC)), and (iv) 171, 193, and 211 Å composite image (three-channel composite (3CC)). We then pixel-wise cluster each image using the *k*-means method. This method is used to automatically cluster a given data set into *k* groups of equal variance (MacQueen 1967). The most commonly used clustering criterion is the sum of squared Euclidian distances (SSD), also known as the within-cluster sum of squares, of each data point to the centroid of the cluster, to which that data point is assigned (Likas et al. 2003). The *k*-means algorithm first randomly selects *k*-cluster centroids, and then iteratively refines these initial cluster centroids by assigning each Euclidian distance to its closest cluster centroid. Then the algorithm updates each cluster centroid value to be the mean of its elements by minimizing the SSD (Wagstaff et al. 2001; Likas et al. 2003).

The number of clusters, the *k* value, for this method is an input parameter. To choose the optimum number of clusters, we used the scree-plot method (Paparrizos & Gravano 2015). In this method, we use $k = 1, 2, 3, \cdots, 10$ and calculate the the sum of squared distances (SSD) for each *k* value. The results show that after the cluster number 3, any further decrease in SSD is very small compared to previous ones, which means that the optimum *k* value to use is 3 (Figure 4). This indicates that there are darker regions, brighter regions, and regions that surround them, which can be assigned to the CHs, active regions, and the quiet Sun.

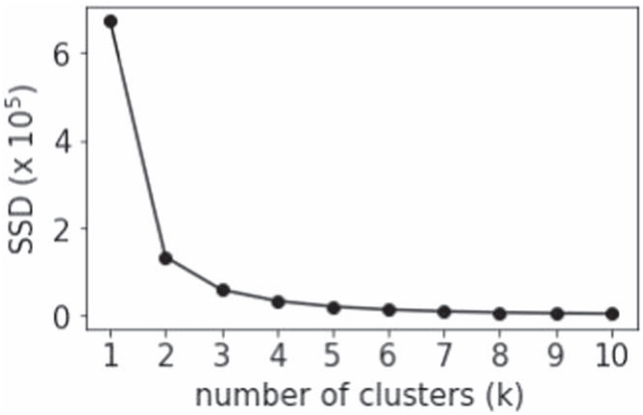

Figure 4. Sum of squared distances (SSD) calculated for each number of clusters, which ranges from 1 to 10 for passband data in 193 Å on 2017 December 8 at 00:00 UT.

The *k*-means method allows us to determine a threshold value for single-channel inputs, a threshold line for two-channel inputs, and a threshold surface for three-channel inputs in a systematic way that enables us to deter from choosing these thresholds arbitrarily. Additionally, this method, when automated, is flexible enough for day-to-day variations in solar images, providing a dynamical response to them.

We calculate segmentation maps for each date using the *k*-means method throughout solar cycle 24. Following that, we convert these maps to binary maps by merging the two clusters that identify brighter regions (active regions) and regions that surround darker and brighter regions (quiet Sun). The reason we did not use a *k* value of 2, is to avoid overestimation of the

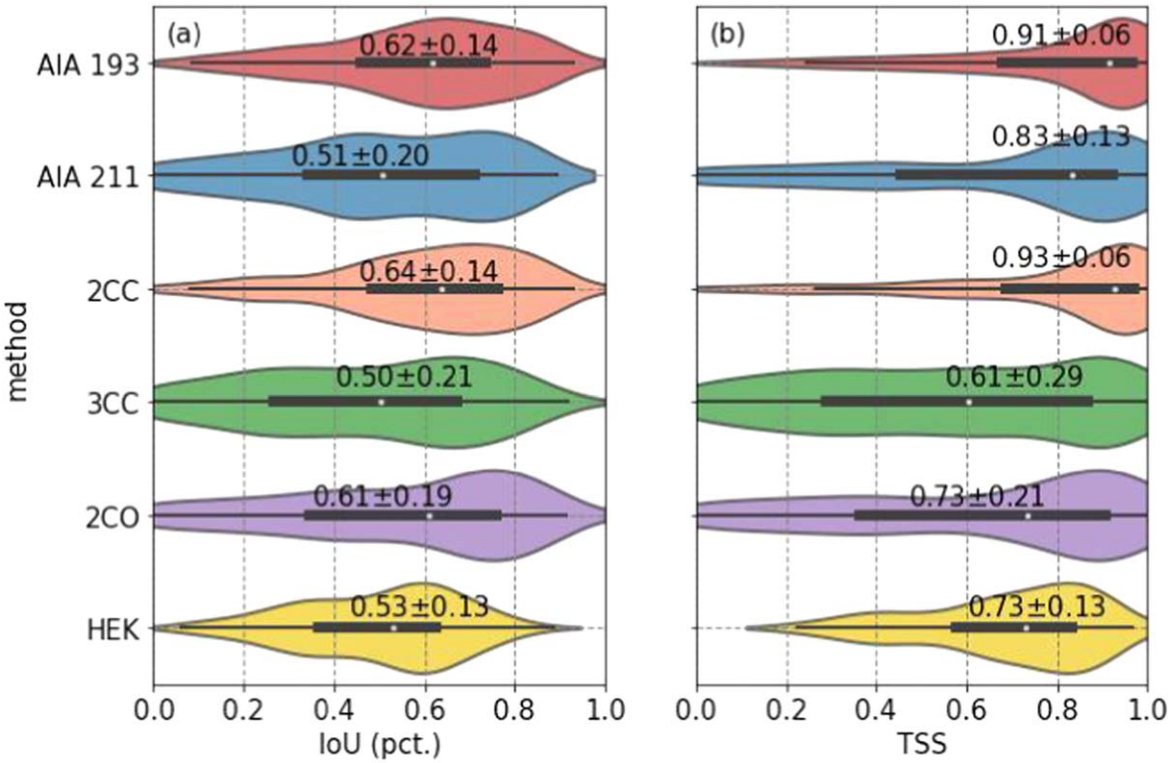

Figure 5. The distributions of the calculated IoU (a) and TSS (b) values between binary maps generated in this study and the CATCH, together with those between the HEK database and the CATCH. The white dots indicate the median value for each distribution. We also show the median values together with median absolute deviation for each evaluation metric in the figure. The red, blue, orange, green, purple, and yellow colors show AIA 193, AIA 211, 2CC, 3CC, 2CO, and HEK binary maps, respectively.

darker pixels on the passband images of the solar disk. We then remove small dotted-like regions using the *morphology* module of the scikit-image package (van der Walt et al. 2014). This method requires two inputs, the smallest allowable object size and connectivity, for which we use 200 and 10 pixels, respectively. We also used morphological closing using a disk-shaped footprint with a radius of 2 pixels to remove smaller holes in identified CHs. The reason for using a smaller footprint is to try to avoid smoothing out larger bright points in identified CHs, which might be related to the coronal bright points (Karachik et al. 2006; Hong et al. 2014; Wyper et al. 2018).

In addition to the four different binary map types generated based on the 193 Å, 211 Å, 2CC, and 3CC, we generated another type of binary map. We generated them based on the overlap between binary maps of the 193 and 211 Å images, which we will refer to as the two-channel overlap (2CO). The 2CO binary maps are created if a pixel is simultaneously identified as a CH pixel in the two binary maps from the 193 and 211 Å images. Those pixels that are not simultaneously identified as a CH are then accepted as non-CH pixels.

3.3. Pixel-wise Evaluation Metrics

To calculate the performances of our binary maps generated by the *k*-means method for each date, we used pixel-wise evaluation metrics. As there will be an imbalance between non-CH and CH pixels in the passband and composite images of the Sun, we use intersection over union (IoU), also known as the Jaccard index (Jaccard 1912), and true skill statistics (TSS; Hanssen & Kuipers 1965) as pixel-wise evaluation metrics. To calculate these metrics, we used binary maps from CATCH. IoU and TSS are calculated based on each confusion matrix for each date using

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{2}$$

where TP, TN, FP, and FN denote the pixel-wise calculated number of true positives, true negatives, false positives, and false negatives, respectively.

The distributions of the IoU values calculated between our and the CATCH binary maps together with those between the HEK and the CATCH binary maps show that the IoU for the HEK CH binary maps has a median value of 0.53 ± 0.13, while our results from the AIA 193 and 2CC show median values of 0.62 ± 0.14 and 0.64 ± 0.14, respectively. This indicates a better overlap of the identified CHs from our method with those generated by CATCH. The other three binary maps from our study, the AIA 212, 3CC, and 2CO, result in IoU values of 0.51 ± 0.20, 0.50 ± 0.21, and 0.61 ± 0.19, respectively (Figure 5(a)).

The median TSS values of the AIA 193 and 2CC are 0.91 ± 0.06 and 0.93 ± 0.06, respectively (Figure 5(b)), while the median TSS value for the HEK is 0.73 ± 0.13. These results indicate that our binary maps generated by AIA 193 and 2CC are more in line with those from CATCH. The AIA 212, 3CC, and 2CO, show median TSS values lower than AIA 193 and 2CC (Figure 5(b)).

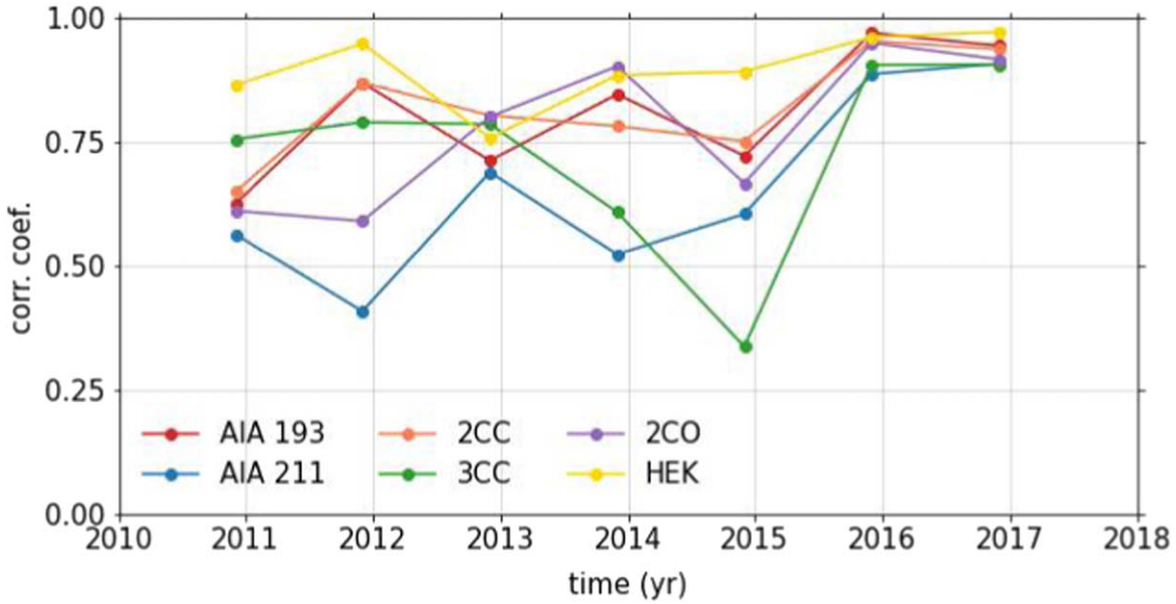

Figure 6. Temporal evolution of the correlation coefficients between total CH areas from our method, HEK against CATCH data through 2010 November and 2016 December, extending through solar cycle 24. Note that the correlations are calculated using data during the last two months of each year (see the text).

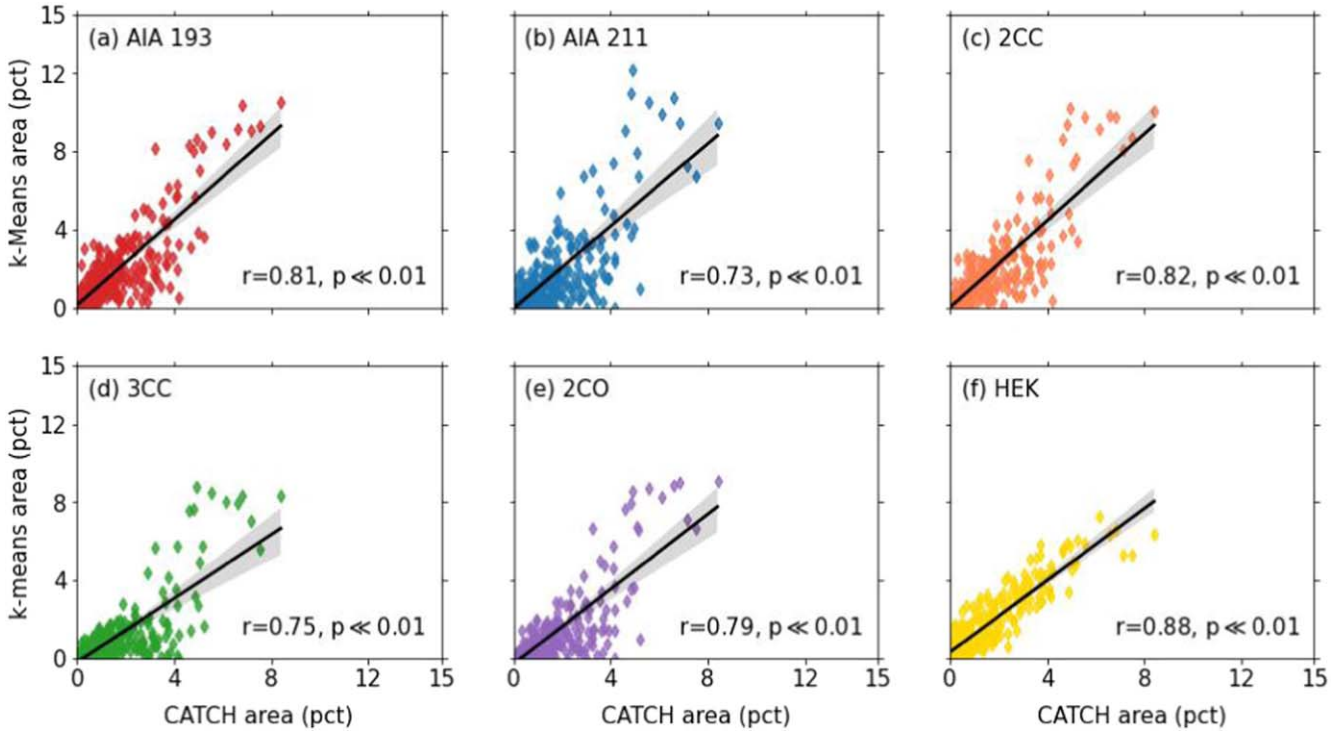

Figure 7. The total percentage areas from this study (a)–(e) and HEK database (f) as a function of the areas from CATCH. The black solid lines show the linear fits, while the shaded areas show uncertainty. We also show the Pearson correlation coefficients and their statistical significances. The color coding is the same in Figure 5.

3.4. Coronal Hole Areas

To further validate our results against the HEK and CATCH results, we calculate the total areas of the CHs on the solar disk in percentage of CH coverage on the solar disk. To achieve this, we first corrected each pixel in our binary maps for projection effects by applying

$$A_i = \frac{A_{i,\mathrm{proj}}}{\cos \alpha_i}, \tag{3}$$

where A_i and α_i denote the corrected pixel area and the heliographic angular distance of each pixel to the center of the solar disk as seen from the AIA/SDO, respectively.

We calculated the Pearson correlation coefficients for each year between results from our study, HEK binary maps, and CATCH (Figure 6). We need to note that we use the last two months of each year to calculate the correlations. Similar to the results obtained for IoU and TSS, AIA 193 and 2CC generally provide higher correlations through the study period.

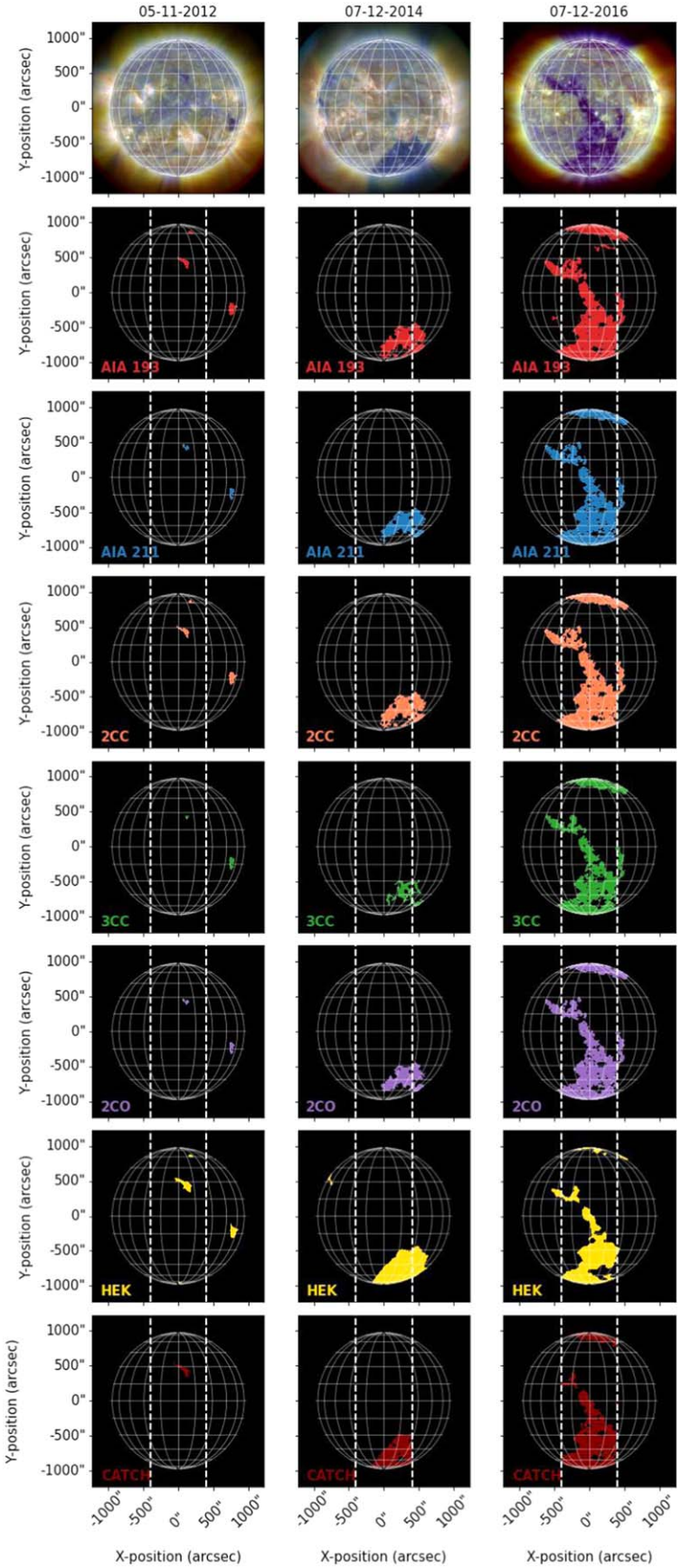

Figure 8. The CH binary maps for 2012 November 5 (top row), 2014 December 7 (middle row), and 2016 December 7 (bottom row) identified from the AIA 193, AIA 211, 2CC, 3CC, 2CO together with binary maps from the HEK and CATCH. The vertical white dashed lines indicate the longitudinal range of [−400, 400] arcsec in helioprojective coordinates. The color coding is the same in Figure 5.

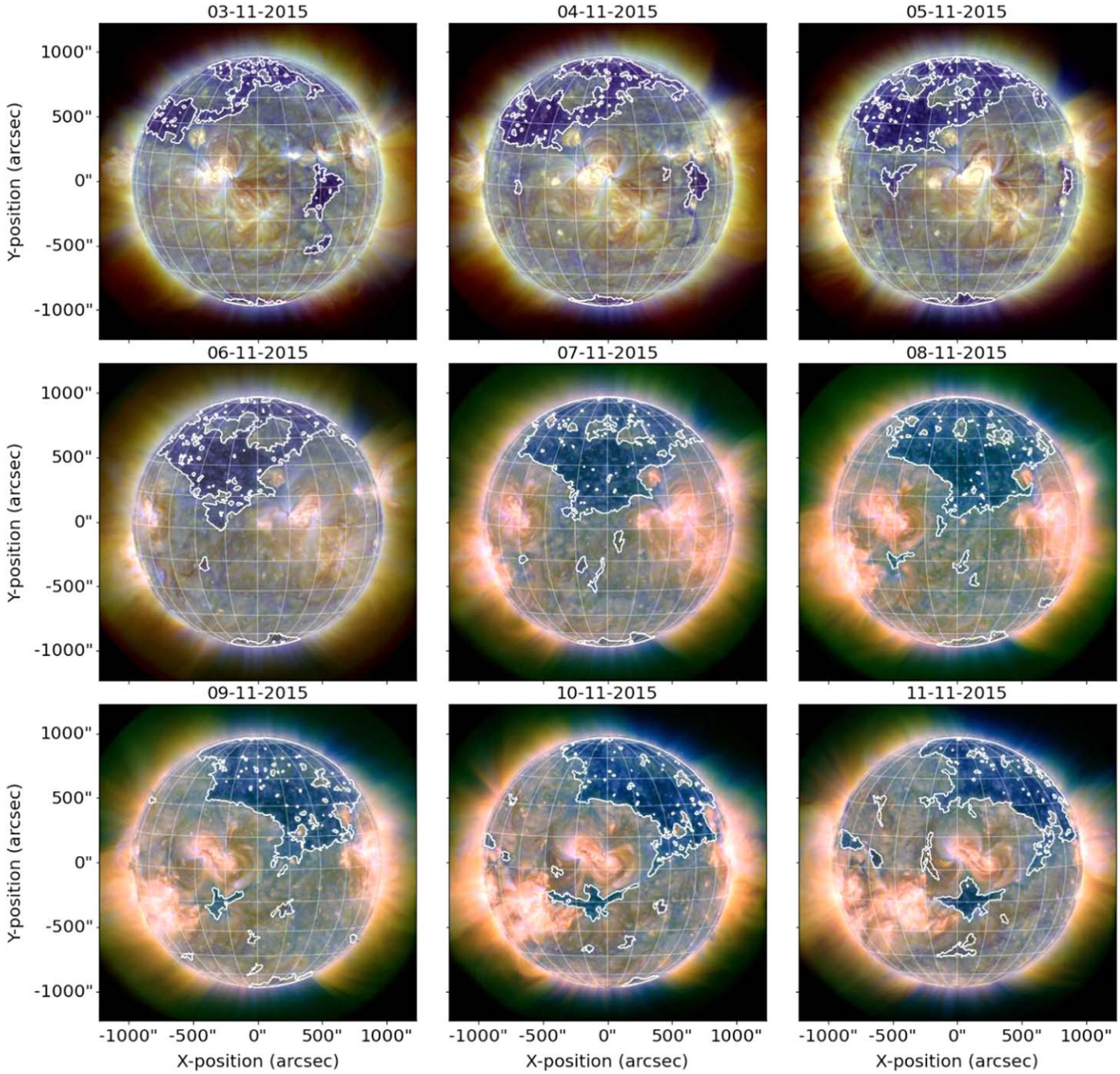

Figure 9. The CH binary maps for a time sequence from 2015 November 3 through 2015 November 11 identified from the 2CC.

Interestingly after 2014, the correlation coefficients calculated for every binary map become similar and evolve in parallel until 2016 (Figure 6).

We also calculated the overall correlations between the binary maps from our study and HEK, and binary maps from CATCH. The highest correlation of 0.88 for the CH areas is observed between the HEK and the CATCH data, while our 2CC gave a correlation coefficient of 0.82, followed closely by AIA 193 that gave a correlation coefficient of 0.81. The correlation coefficients for the 2CO, 3CC, and AIA 212 are 0.79, 0.75, and 0.73 respectively (Figure 7).

3.5. Comparison of the CH Binary Maps

We then select three dates that represent different phases of solar cycle 24 to compare the CH binary maps. These dates are (i) 2012 November 5 on the inclining phase before the cycle maximum, (ii) 2014 December 7 right after the solar cycle maximum, and (iii) 2016 December 7 on the declining phase of solar cycle 24 (Figure 8).

On the inclining phase of solar cycle 24, on 2012 November 5, our method identifies smaller CHs. The results from the AIA 193, 3CC, and 2CO are observed to be more in line with those from the CATCH, where there is only one CH at [0, 500]

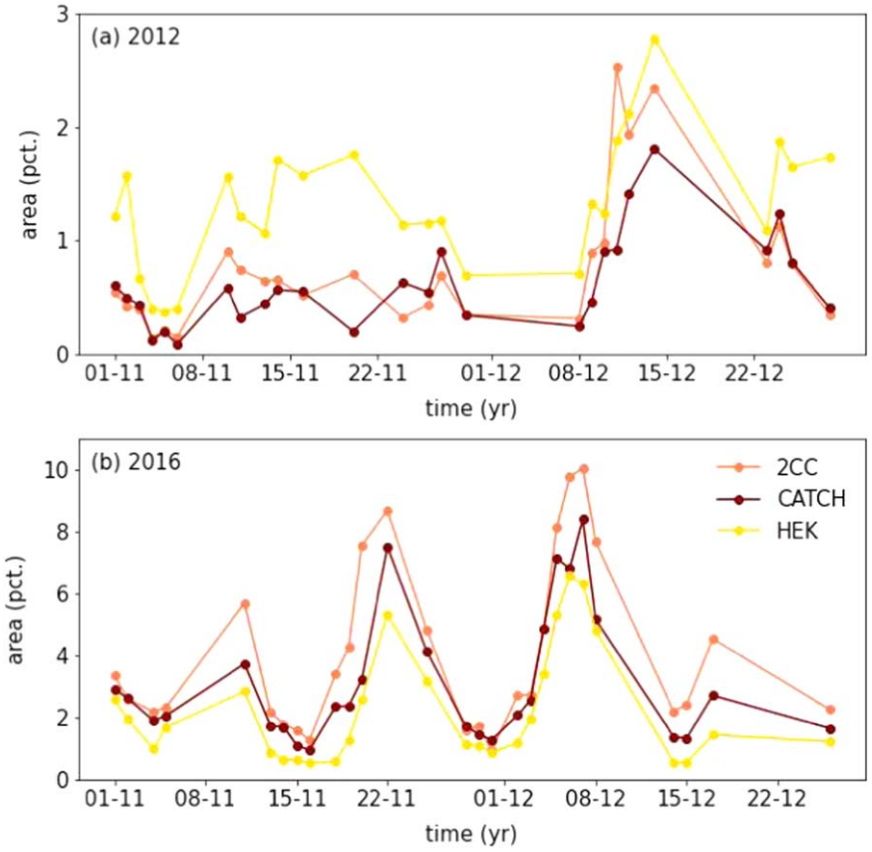

Figure 10. The CH areas during the last two months of 2012 (a) and 2016 (b). The coral, gold, and maroon lines represent 2CC, HEK, and CATCH data, respectively.

arcsec in helioprojective coordinates. The results from the AIA 211 and the 2CC, on the other hand, are more in line with those from the HEK database (the top row of Figure 8). On 2014 December 7, a few months after the cycle maximum, the binary maps from the AIA 193, the 3CC, and the 2CO show similar CH coverage on the solar disk to the CATCH within the longitudinal range of [−400, 400] arcsec. All of the CH binary maps from our method, except for the 3CC, are similar to the CHs from the HEK showing a small coronal hole near [−750, 500] arcsec (the middle row of Figure 8). On the declining phase of solar cycle 24, on 2016 December 7, the CH areas identified using the AIA 193, the 2CC, and the 3CC are in line with those from the HEK database and CATCH. On this date, the total CH area coverage also reaches its maximum, where it extends from the southern solar pole to the solar equator (the bottom row of Figure 8).

To evaluate the consistency of our results, we plotted the detected CHs using 2CC on the dates from 2015 November 3 through 2015 November 11 (Figure 9). The temporal evolution of the detected CHs close to the solar equator is consistent with the solar rotation. Formation and evolution of a new CH, again close to the solar equator, starting from the 6th of November through 11th of November can also be observed. In addition, temporal evolution of the large CH on the northern solar hemisphere is also consistent in each date (Figure 9).

To further investigate the consistency, we checked the day-to-day temporal evolution of the areas during 2012 and 2016 (Figure 10). Note that the areas are calculated for the last two months of each year. In 2012, there is a general good agreement between our 2CC, CATCH, and HEK CHs especially during December, whereas in November, the HEK CH areas are larger compared to our 2CC and the CATCH (Figure 10(a)). During 2016, on the other hand, CH areas from the three sources covary with some small differences in amplitudes (Figure 10(b)).

4. Discussion and Conclusions

CHs are the source regions of the steady fast solar winds, which result in corotating interaction region-driven storms, the so-called HILDCAA events (Tsurutani & Gonzalez 1987). In comparison to their surroundings, CH have lower plasma densities and temperatures and therefore they have the lowest emissions in the UV and X-ray wavelength range. This physical feature makes them appear as darker regions in passband images of the Sun taken in these wavelengths. CHs are also known to have very complex magnetic structures extending from the photosphere to the corona (Heinemann et al. 2018, 2021), where the open magnetic field lines extend into the interplanetary medium. They also show solar cycle dependence.

There are several methods to identify CHs on the solar images taken by AIA/SDO and EIT/SOHO based on histograms (Krista & Gallagher 2009), multithermal intensity segmentation (Garton et al. 2018), and intensity threshold, which is modulated by the intensity gradient of a CH (Heinemann et al. 2019). Recently, unsupervised and supervised ML methods have been used to detect CHs using single- or multichannel passband data from the AIA/SDO (Verbeeck et al. 2014; Illarionov & Tlatov 2018; Jarolim et al. 2021). The supervised ML methods mainly rely on the CNNs for image segmentation. These methods, however, require a reliable

training data set that consists of CH polygons detected either by an observer or by an unsupervised method.

In our study, to identify the CHs we used a simple clustering algorithm, *k*-means, to pixel-wise cluster the passband images of the Sun taken in 171, 193, and 211 Å by the AIA/SDO covering the time period between 2010 November and 2016 December. In addition to using a single-channel approach, we used different combinations of these channels. To detect the lower- and upper-threshold values, we fitted bimodal Gaussians to the probability densities of intensities for each channel on each date. We then calculated the thresholds based on the mean and standard deviation of the local maximum at higher intensities. To cluster the passband images, we used the *k*-means method, where the optimum number of clusters, 3, is calculated based on the scree plot. The *k*-means method, together with pre- and postprocessing steps enabled us to build an automated flexible approach that dynamically responds to day-to-day variations in solar images. As a result we obtained five different binary maps for each identified CHs, that are (i) AIA 193, (ii) AIA 211, (iii) 2CC, (iv) 3CC, and (v) 2CO. We then calculated pixel-wise evaluation metrics based on CH binary maps from CATCH and compared our results with them as well as those from the HEK database. Following that, we calculated the total percentage area identified as a CH per date, after correcting the binary maps for the projection effects.

Our results show that the 2CC, a composite image using only 193 and 211 Å passband images, provides the best results that are closely followed by results from AIA 193. The median IoU and TSS values for the 2CC are 0.64 ± 0.14 and 0.93 ± 0.06, respectively, while they are 0.62 ± 0.14 and 0.91 ± 0.06 for the AIA 193. Our results show higher similarity to CATCH results than the HEK database ($\mathrm{IoU} = 0.53 \pm 0.13$ and $\mathrm{TSS} = 0.73 \pm 0.13$). Our results provided better overlap with the CATCH data than those obtained by the CHRONNOS method (Jarolim et al. 2021) for the same period, which provided mean IoU and TSS values of 0.63 and 0.81, respectively. This method uses all of the seven channels from the AIA/SDO and line-of-sight magnetograms from the HMI/SDO in progressively growing CNNs (Jarolim et al. 2021). Even though our results from AIA 193 and 2CC also provide high overall correlations, they are still lower than the correlation coefficient of 0.88 between the HEK binary maps and CATCH.

We also showed the consistency of our results, especially from the 2CC method, when the formation and temporal evolution of the CHs are considered. Our method was able to identify and track the CHs from November 3 through November 11 for nine consecutive days. Additionally, temporal variations of CH areas from our method follows the trends that are observed in the CATCH and HEK CH areas.

To investigate the effects of the chosen lower- and upper-threshold values, we also calculated the same evaluation metrics and areas for the threshold ranges of $\mu \pm 3\sigma$ and $\mu \pm 5\sigma$, as well as for cases where we do not apply any thresholding at all. Similarly, we calculated the thresholds based on the bimodal Gaussian fit and the mean and standard deviation of the local maximum at the higher intensities. However, using different thresholds, and also not using any thresholds, provided lower evaluation metrics as well as correlation coefficients of the total areas.

Interestingly enough, our results show significant discrepancies between the identified CHs using our method, HEK, and CATCH when we look at the temporal variations in the correlation coefficients calculated for the total areas. Recently, some steps have been taken to create a reliable database where there is a consensus about the CH boundaries and their uncertainties are being discussed (Linker et al. 2021; Reiss et al. 2021).

In conclusion, as an unsupervised ML method, using the *k*-means clustering provides better results with those from complex methods, such as CNNs. One of the most important steps in this method is the preprocessing of the data and the choice of the lower- and upper-threshold values in a more systematic way, which then can lead to automation of the CH detection at any given date or a date range. More importantly, our study shows that there is a need for a CH database where a consensus about the CH boundaries is reached by observers independently, and which can be used as the "ground truth," when using a supervised method or just to evaluate the goodness of the models.

This research is supported by the Helmholtz Imaging Platform, Solar Image-based Modeling (SIM) ZT-I-PF4-016.

ORCID iDs

Fadil Inceoglu https://orcid.org/0000-0003-4726-3994
Yuri Y. Shprits https://orcid.org/0000-0002-9625-0834

References

Barnes, W., Cheung, M., Bobra, M., et al. 2020a, aiapy: A Python Package for Analyzing Solar EUV Image Data from AIA v0.3.1, Zenodo, doi:10.5281/zenodo.4274931
Barnes, W. T., Cheung, M. C. M., Bobra, M. G., et al. 2020b, JOSS, 5, 2801
Cranmer, S. R. 2009, LRSP, 6, 3
Delaboudinière, J. P., Artzner, G. E., Brunaud, J., et al. 1995, SoPh, 162, 291
Delouille, V., Hofmeister, S. J., Reiss, M. A., et al. 2018, in Machine Learning Techniques for Space Weather, ed. E. Camporeale, S. Wing, & J. R. Johnson (Amsterdam: Elsevier), 365
Eastwood, J. P., Biffis, E., Hapgood, M. A., et al. 2017, RiskA, 37, 206
Garton, T. M., Gallagher, P. T., & Murray, S. A. 2018, JSWSC, 8, A02
Hanssen, A., & Kuipers, W. 1965, Kon. Neder. Meteor. Inst. Meded. Verhand, 81, 15
Harvey, K. L., & Recely, F. 2002, SoPh, 211, 31
Heinemann, S. G., Hofmeister, S. J., Veronig, A. M., & Temmer, M. 2018, ApJ, 863, 29
Heinemann, S. G., Temmer, M., Heinemann, N., et al. 2019, SoPh, 294, 144
Heinemann, S. G., Temmer, M., Hofmeister, S. J., et al. 2021, SoPh, 296, 141
Hewins, I. M., Gibson, S. E., Webb, D. F., et al. 2020, SoPh, 295, 161
Hong, J., Jiang, Y., Yang, J., et al. 2014, ApJ, 796, 73
Hurlburt, N., Cheung, M., Schrijver, C., et al. 2012, SoPh, 275, 67
Illarionov, E. A., & Tlatov, A. G. 2018, MNRAS, 481, 5014
Jaccard, P. 1912, New Phytol., 11, 37
Jarolim, R., Veronig, A. M., Hofmeister, S., et al. 2021, A&A, 652, A13
Karachik, N., Pevtsov, A. A., & Sattarov, I. 2006, ApJ, 642, 562
Krista, L. D., & Gallagher, P. T. 2009, SoPh, 256, 87
Lecun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, SoPh, 275, 17
Likas, A., Vlassis, N., & Verbeek, J. 2003, PatRe, 36, 451
Linker, J. A., Heinemann, S. G., Temmer, M., et al. 2021, ApJ, 918, 21
MacQueen, J. 1967, in Proc. 5th Berkeley Symp.on Mathematical Statistics and Probability, 1: Statistics, ed. L. M. Le Cam & J. Neyman (Berkeley, CA: Univ. California Press), 281
Marsch, E. 2006, LRSP, 3, 1
Mumford, S. J., Freij, N., Christe, S., et al. 2021, SunPy v3.0.3, Zenodo, doi:10.5281/zenodo.5751998
Paparrizos, J., & Gravano, L. 2015, in Proc. 2015 ACM SIGMOD Int. Conf. on Management of Data, SIGMOD '15 (New York: Association for Computing Machinery), 1855
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3
Reiss, M. A., Muglach, K., Möstl, C., et al. 2021, ApJ, 913, 28

Ronneberger, O., Fischer, P., & Brox, T. 2015, in Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, ed. N. Navab et al. (Berlin: Springer), 234
Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, SoPh, 275, 207
Schmidhuber, J. 2014, NN, 61, 85
Schwabe, H. 1844, AN, 21, 233
Schwenn, R. 2006, LRSP, 3, 2
The SunPy Community, Barnes, W. T., Bobra, M. G., et al. 2020, ApJ, 890, 68
Tsurutani, B. T., & Gonzalez, W. D. 1987, P&SS, 35, 405
van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., et al. 2014, PeerJ, 2, e453
Verbeeck, C., Delouille, V., Mampaey, B., & De Visscher, R. 2014, A&A, 561, A29
Wagstaff, K., Cardie, C., Rogers, S., et al. 2001, in Proc. of the 18th Int. Conf. on Machine Learning, ICML '01 (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 577
Wilcox, J. M. 1968, SSRv, 8, 258
Wyper, P. F., DeVore, C. R., Karpen, J. T., Antiochos, S. K., & Yeates, A. R. 2018, ApJ, 864, 165