The Astrophysical Journal, 889:153 (9pp), 2020 February 1 https://doi.org/10.3847/1538-4357/ab5a7a

Objectively Determining States of the Solar Wind Using Machine Learning

D. Aaron Roberts[1], Homa Karimabadi[2], Tamara Sipes[3], Yuan-Kuen Ko[4], and Susan Lepri[5]
[1] Heliophysics Division, NASA Goddard Space Flight Center, Code 672, Greenbelt, MD 20771, USA; aaron.roberts@nasa.gov
[2] Analytics Ventures, 6450 Lusk Blvd, Suite E208, San Diego, CA 92121, USA
[3] Optum, 6195 Lusk Blvd, Suite 120, San Diego, CA 92121, USA
[4] Space Science Division, Naval Research Laboratory, Washington, DC 20375, USA
[5] Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI 48109, USA

Received 2019 August 20; revised 2019 November 20; accepted 2019 November 20; published 2020 February 3

Abstract

Conclusively determining the states of the solar wind will aid in tracing the origins of those states to the Sun, and in the process help to find the wind's origin and acceleration mechanism(s). Prior studies have characterized the various states of the wind, making lists that are only partially based on objective criteria; different approaches obtain substantially different results. To uncover the unbiased states of the solar wind, we use "*k*-means clustering"—an unsupervised machine learning method—including constructed multipoint variables. The method allows exploration of different descriptive state variables and numbers of fundamental states (clusters). We show that the clusters reveal structures similar to those found by more ad hoc means, including coronal hole wind, interplanetary coronal mass ejections, "slow wind" (better: noncoronal hole flow), "pseudostreamers," and stream interaction regions, but with differences that should be useful in refining our previous ideas. These results demonstrate the viability of the approach and warrant further study to understand the origin of remaining discrepancies. Complexity in *k*-means characterization of the wind may ultimately point to complexity at the source; studies closer to the Sun with *Parker Solar Probe* will help. We confirm the utility of a set of variables that can serve as a proxy for composition measurements. This proxy permits studies at high time resolution and where composition is not available. This and our recently developed unsupervised multivariate clustering technique are expected to be beneficial in the automated identification of structures and events in a variety of studies.

Unified Astronomy Thesaurus concepts: Interplanetary physics (827); Solar wind (1534); Interplanetary turbulence (830); Interplanetary magnetic fields (824); Solar coronal mass ejections (310)

1. Introduction

The ubiquitous, continuous flow of fully ionized particles from the Sun—the solar wind—is much more variable than the Earth's atmosphere. As with our local atmosphere, we would like to have clear and useful characterizations of the variability, things equivalent to "hot, dry, and sunny" versus "cool and foggy," versus "category 4 hurricane." For the solar case, such characterizations could be useful in determining the origins and effects of different flows (see, e.g., Zhao et al. 2017, for an approach starting with solar states). Time series of even a simple set of, say, speed, density, and magnetic field variables from the solar wind invite the eye to informally categorize different "states," most obviously fast and slow, but with further examination including various transient events, "sector boundaries," and more subtle regions such as shocked plasma or places where electron flows are anomalous. (A detailed overview of studies of solar wind states is provided by Borovsky et al. 2019; below we discuss the most relevant studies for the current work.)

In a general sense, many of these states are understood: magnetically open, X-ray-weak areas of the solar corona ("holes") are the clear origin of most very fast wind, and slow wind likely comes from near the edges of such holes and from the "streamer stalks" so prominent in deep-exposure eclipse photographs or coronal images from spacecraft. Magnetic arcades known as filaments or prominences, depending on whether they are viewed against the solar surface or off to the side (on the limb), sometimes give rise to transient coronal mass ejections (CMEs) that often become interplanetary flux ropes (a type of interplanetary CME, or ICME) that are observed in situ by spacecraft. It has proven difficult to definitively characterize the various types of wind from in situ measurements, such that various schemes for classification (e.g., Jian et al. 2006; Zhao et al. 2009; Richardson & Cane 2010; Xu & Borovsky 2015) can disagree a large fraction of the time (Neugebauer et al. 2016). There may well be usefully distinct types of ICMEs or even fast coronal hole flows.

All current schemes for classifying solar wind states ultimately depend on (1) a predetermined set of categories, and (2) an often-subjective evaluation of the presence or absence of those categories based on characteristics such as flow speed, the presence of smooth magnetic rotations, and the properties of charge states of particular ions. There are a number of clear-cut cases, such as the large magnetic field rotation associated with "magnetic-cloud" ICMEs or the seemingly unperturbed flow in the middle of a high-speed stream. Such relatively obvious cases can be the origin of analytical expressions (e.g., Zhao et al. 2009; Xu & Borovsky 2015) or can provide training sets for "supervised" methods (see, in this context, Camporeale et al. 2017) that can then be used to categorize the rest of the events. These approaches can be helpful for performing statistical studies, but the fact that different investigators disagree on events indicates either that we have not determined a fully convincing method of characterization or perhaps that there are no truly reliable general categories to be found. For a general discussion of the larger context and use of "machine learning" methods for the analysis of space physics data, see McGranaghan et al. (2017).

2. Method and Data

2.1. The k-means Algorithm

Much of the previous work in this area has relied on finding clusters in scatter plots of various variables. This can be done visually in two dimensions, but this method rapidly fails when going to higher dimensions. Moreover, when a random set of states of the wind is viewed in such (low-dimensional) plots, it never separates out cleanly into different clumps. Even the "ideal" cases are not completely disentangled. This suggests the possibility that what is needed is a higher dimensional approach, such that additional dimensions sort out the confusion. It turns out that the problem of finding clusters in high dimensions was largely solved many years ago, and there are now a number of clustering schemes that use different criteria for grouping. One of the most intuitive and easy to calculate is the "*k*-means" algorithm, developed over 50 years ago and now implemented in various software packages. (The term and a standard exposition are given by MacQueen (1967), but the history goes back earlier[6].) In the implementation used here, the method starts by assuming that there are k clusters, each centered on a point in an N-dimensional state space. The centroids are initially chosen randomly, and the distances from the k centroids to all points are determined. The points closest to each centroid are then used to calculate new centroids, and this process is iterated until very little change occurs. The result is k clumps that are taken to characterize interesting "natural" states of the system described by the variables. In practice, ten or so iterations are used to obtain adequate convergence. We have chosen the *k*-means method over others as an initial attempt to do meaningful clustering in a way that is very efficient, allowing many trials with different variables and parameters, but that also captures very directly a physicist's notion of "natural states." We have tried some less efficient variants of the basic procedure, such as "Gaussian mixture models"[7] with essentially similar results.

The *k*-means procedure is equivalent to minimizing the total variance of the states from the centroids. Mathematically, given M states $\boldsymbol{x}_i$ in an N-dimensional state space, the "objective function" to be minimized is

$$J = \sum_{i=1}^{M}\sum_{j=1}^{k} w_{ij}||\boldsymbol{x}_i - \mu_j||^2 \tag{1}$$

where μ_j is the N-vector centroid of the jth cluster, w_{ij} equals one if point i is a member of cluster j and zero otherwise, and the norm is taken here to be the standard Euclidean distance in the N-dimensional state space. Other norms are possible, and in fact the IDL-provided code for *k*-means uses the "Manhattan" norm of the sum of absolute values of component differences instead of the sum of squares. The alternating steps of the minimization choose w_{ij} by determining the points nearest to the current centroids, and then finding a new set of values for μ_j by averaging over the new w_{ij} sets. Given IDL's nonstandard norm and other nonstandard features (e.g., the mean of the cluster values does not end up being exactly the reported centroid), we wrote our own version of the algorithm (still in IDL), exactly implementing the process described above. We found that some initial conditions selected essentially isolated points that never acquired many other points, so we rejected initial centroid positions in which one of the initial clusters had fewer than ten points. It may be that some of the nearly isolated points are interesting outliers, but we leave this issue to future work.

If the natural states (clusters) occur in well-separated convex regions, this method works quickly and well. Possible pitfalls are finding a local minimum in the space but missing the global minimum; not correctly characterizing the shapes of the clusters; and having the "wrong" value of k. Multiple runs with different ks and initial centroids can help to sort out the first and third of these problems, and visual examination of 2D projections of the distribution can give some reassurance about the second. Here we found that 30 trials were enough to arrive at a very slowly changing J, and results presented here are based on this; there is never a complete convergence in complex data such as these, but we plan to study more carefully how the results evolve with decreasing J. Changing k values will be dealt with below. While there are a variety of clustering techniques (Pedregosa et al. 2011),[8] *k*-means is a widely used technique due to its simplicity and general utility. This technique is not sensitive to any temporal correlations of points in the time series data. Note that many of the recent advances in artificial intelligence, such as convolutional neural nets, have centered around "supervised" methods that require large numbers of labeled data sets. Such techniques are clearly not directly applicable to the solar wind categorization problem, where the objective is to find the proper "labels" for the different states of the solar wind. However, a recently developed unsupervised technique leverages advances in artificial intelligence and takes into account any underlying temporal correlations in time series data (Madiraju et al. 2018). We will explore application of this technique to the solar wind classification problem in a separate publication.

2.2. Data Set Used

The data examined here are essentially the same as those used in the study of Neugebauer et al. (2016), namely magnetic field (from the MAG instrument), plasma parameters (density, velocity, alpha-to-proton ratio; SWEPAM), and ion composition (SWICS), all from the *Advanced Composition Explorer* (*ACE*) spacecraft upstream of the Earth at the L1 point. Hourly binned quantities were determined for each instrument so that they could be compared directly. The time period analyzed is 2002 November through 2004 May (a total of 20,448 points), when there was sufficient solar activity to produce many ICMEs. All quantities, z, were normalized to the range from 0 to 1 by taking $z_n = (z - \min(z))/(\max(z) - \min(z))$. This is necessary to make density (say) with a range below 20 cm^{-3} contribute as much to distances in the state space as, for example, the solar wind speeds of 250–1000 km s^{-1}. Future work will examine the effects of weighting the variables differently using multipliers, but here all weights of the

---

[6] See https://en.wikipedia.org/wiki/k-means_clustering.

[7] See https://scikit-learn.org/stable/modules/mixture.html#mixture.

[8] See scikit-learn https://scikit-learn.org/stable/modules/clustering.html#clustering.

normalized variables are taken to be equal. In some cases, explicitly indicated, the log of the variable is used to spread the distribution of a variable such as density that tends to be concentrated at small values. Often a normalization of zero mean and unit standard deviation is used for the variables, but this was not found to give as clear results in this case, perhaps due to the highly non-Gaussian distributions of the variables.

2.3. “Nonlocal” Variables

Given the normalized variables, any set or combination of them can be used to represent the solar wind state. In addition to a number of direct measurements, we use the time correlations known as the “cross-helicity”

$$\sigma_c = \frac{2\langle \delta\boldsymbol{b} \cdot \delta\boldsymbol{v}\rangle}{\langle(\delta\boldsymbol{b})^2\rangle + \langle(\delta\boldsymbol{v})^2\rangle} \tag{2}$$

and the “residual energy”

$$\sigma_r = \frac{\langle(\delta\boldsymbol{v})^2\rangle - \langle(\delta\boldsymbol{b})^2\rangle}{\langle(\delta\boldsymbol{b})^2\rangle + \langle(\delta\boldsymbol{v})^2\rangle} \tag{3}$$

where the brackets $\langle\rangle$ represent averages, here taken over three data points, although the main results are essentially the same with, e.g., seven-point averages. The δs indicate that a running mean has been subtracted out over the averaging interval. Here, $\boldsymbol{v}$ is the wind velocity and $\boldsymbol{b}$ is the magnetic field in “Alfvén speed units, ” which allows $\boldsymbol{b}$ and $\boldsymbol{v}$ to be compared directly:

$$\boldsymbol{b} = \frac{21.8\boldsymbol{B}/\mathrm{nT}}{\sqrt{n/\mathrm{cm}^{-3}}} \tag{4}$$

for a magnetic field $\boldsymbol{B}$ and ion density n. These two quantities have proven quite useful in the study of solar wind turbulence (e.g., Bavassano et al. 1998; Wicks et al. 2013), and here we will find that they are useful for categorizing wind states. Note that this implies that the “state” of the system depends not just on values of quantities at a given time, but also on the variables nearby in time (shear, currents, etc.). The cross-helicity is a measure of how purely the mode of the wind fluctuations is that of an Alfvén wave ($\sigma_c = \pm 1$). Such waves strongly tend to propagate outward from the Sun, and the sign is determined by the direction of the mean magnetic field. Thus, σ_c provides an indication of whether the flow is from the southern or northern hemisphere of the Sun.

2.4. Prior ICME Identifications

The validity of the results cannot be determined by simple means, since we are not assuming any a priori categories. As guides to significance, we use two sets of ICME identifications (Jian et al. 2006; Richardson & Cane 2010) and the analytical formulae of Zhao et al. (2009). The latter workers identify three states: coronal hole flows (nominally “fast wind”), noncoronal hole flows (nominally “slow wind”), and transients (ICMEs). They use the value of the ratio of the number densities of two charge states of oxygen, O^{+7} to O^{+6} (“o7to6”). Coronal hole flows have $\mathrm{o7to6} < 0.145$. The ICMEs have

$$\mathrm{o7to6} > 6.008\exp(-0.00578 v_p) \tag{5}$$

where v_p is the wind (proton) speed, and anything else is noncoronal hole flow. A primary issue addressed by this set of criteria is that it is not a priori obvious where the boundary is between “fast” wind and “slow” wind; in fact, at times “slow” wind has most of the usual characteristics of “fast” wind, so this division does not seem accurate (see, e.g., Marsch et al. 1981; Roberts et al. 1987). The use of the charge states is based on the idea that these states tell us what the temperature of the solar corona was where the ratios of charged states were formed, in the collisional region close to the Sun, and thus these states provide a physically motivated choice for wind types. The ICME identifications come from a careful study of clear cases based on other criteria (see Zhao et al. 2009).

3. Results

3.1. Clusters Using Charge State Information

We apply the k-means method to the above *ACE* data set, initially choosing eight variables based on the relevance of these variables in previous studies. To capture ICMEs, we include charge state information. In the first example of the k-means process shown here, eight variables describe the state (proton density, n_p; proton speed, V_p; σ_c, σ_r, o7to6, the average ionization of iron, the ratio of charged iron to charged oxygen density (fetoo), and magnetic field strength B). We show $k = 8$ here because it provides many connections to previously studied states; there is nothing special about the equality of k and the number of variables. Higher values of k can divide clusters and yield, for example, two sets of ICMEs, and lower k values lump clusters together. In the k-means context, the choice of variables and number of clusters cannot be automated, and it is the primary subjective aspect of this method. However, this subjectivity is quite different from that involved in the point-by-point determination of wind states.

The results of the present method for the case of eight variables and eight states are shown in Figures 1 and 2, which are based on an illustrative subset (2003 May–August) of the entire time range. The top panel in Figure 1 shows the proton speed, V_p, at the top, normalized x (black) and y (red) components of the magnetic fields in Geocentric Solar Ecliptic coordinates at the bottom, and traces in the middle that show high values for the presence of ICMEs as determined by Jian et al. (2006) (blue, and lower max) and Richardson & Cane (2010) (black). (An immediate note is the substantial differences between the two sets of ICMEs, indicating the need for a more objective classification of these states; see Neugebauer et al. 2016.) The colors of the V_p trace show states as identified by k-means clustering. The bottom panel in Figure 1 is a plot of o7to6 colored by the same states as above. The analytic classification is shown by a dashed horizontal line at $\mathrm{o7to6} = 0.145$, which is the maximum for coronal hole wind, and by a brown dotted trace that provides the lower limit for ICME identification. Noncoronal hole wind is between the dashed and dotted lines. The dotted line is not shown when it goes below 0.145.

As is standard with clustering methods, the clusters do not tell us what they are in physical terms. Here we use prior expert identifications of physical regions to make associations. One qualitative test of the validity of the results is the ease with which this can be done. To begin, we immediately see the recurrence of the 27 day solar rotation in the plot of proton

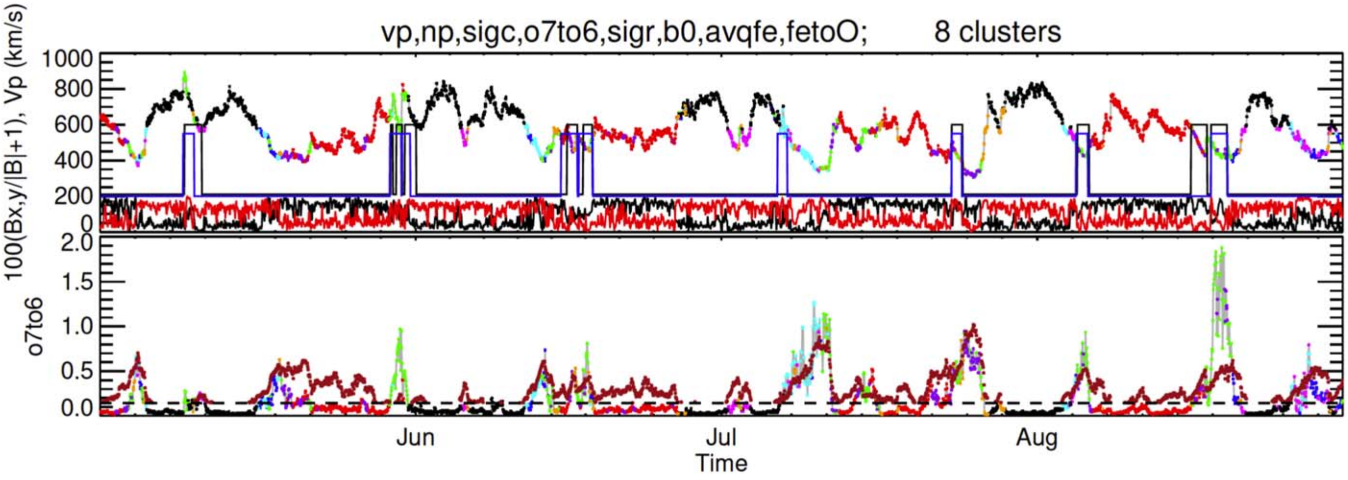

Figure 1. A $k = 8$ set of clusters for four months of 2003 seen as colors superimposed on the plasma speed. Top panel: speed colored by cluster (see Table 1), along with expertly determined ICMEs as deviations from a constant (Richardson: taller black, Jian: shorter blue), and with B_x and B_y as black and red traces below to indicate the magnetic polarity of the solar wind. Lower panel: plot of o7to6 ratio along with a horizontal line showing an empirical upper limit on coronal hole flows and brown dots from Equation (5) showing a lower limit for ICME identification (Zhao et al. 2009).

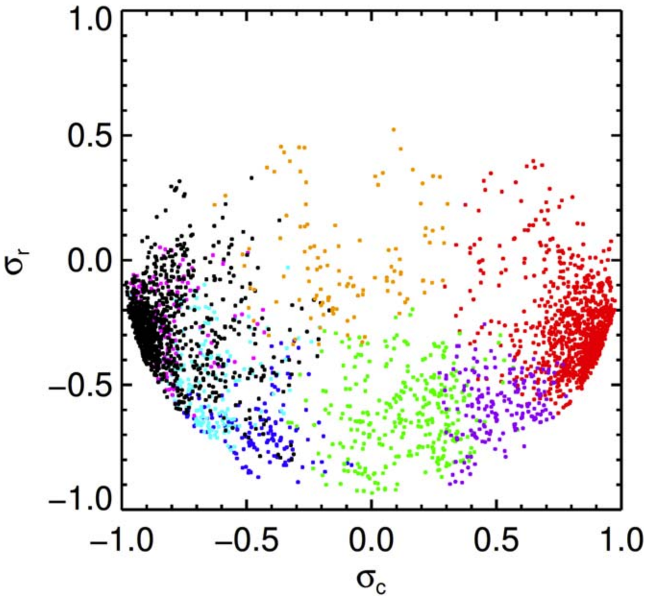

Figure 2. The clusters from Figure 1 seen in the σ_c–σ_r projection of the state space.

speed, V_p, at the top of Figure 1 of particular colors (states), especially black and red. The sector boundary structure, determined by whether the magnetic field is coming from the southern or northern pole of the Sun, is also apparent in, for example, the alternation of black and red as the spacecraft crosses the sector boundary as indicated by the flip in sign of the field in the black and red traces below the main plot in Figure 1 (top). Note that the red sector has systematically slow flow speeds, speeds that would often be termed "slow wind," but both k-means and the analytic formulae agree (the latter having values in the lower panel of Figure 1 that are below the dotted line threshold of 0.145) that these winds are in the same class as the typical fast winds. Figure 2 shows that the black and red states differ in the sign of σ_c, meaning that both contain outward traveling Alfvén waves but in opposite sectors. The input set of variables included wind speed, but this variable did not dominate the choice of category. When the average features of the red and black labeled flows are examined (see Table 1), they are otherwise found to be very similar. Nearly all the points identified by the analytic method as coronal hole flow are either black or red, and almost none of the latter points appear above the 0.145 line.

ICMEs are found to agree significantly with Jian, Richardson, and SWICS lists and criteria, with some interesting differences. The k-means method does not tell us which states are ICMEs, but looking at the first two sets of independently identified ICMEs (nearly all in May) shows that the green state is the natural choice. The first case shows the green agreeing best with the narrower Jian identification, but with a green point slightly later that is still within the Richardson identification. The Jian portion of the interval clearly meets the analytic criterion, but the rest of the interval is marginal, with values nearly equal to the selection criterion. This case is typical of ICME identifications by the various means used here. Detailed investigation of many cases will be needed to see whether k-means can be regarded as more reliable than other methods, but it does provide a rapid, unbiased view.

Some cases where only one of the two expert sets shows a "hit" do not show up in the k-means test (see early July, where the formula marginally predicts an ICME, but k-means does not). There are some identifications that agree with the formula, but not with the lists (around July 10 and 15, for example). One case identifies an ICME in both lists, but in neither k-means nor the formula (the first of the pair in mid-June; the region looks to be complex). Consistent with other work in expert identification, the "green state" of the solar wind has a large average Fe charge state as well as a large value of the field magnitude, B_0. The low value of σ_c is consistent with having typical ICMEs connected at both ends to the Sun, and thus having waves propagating in both directions along the field lines.

What appear to be interaction regions between fast and slow flow (orange) are detected mainly via compression (high density values); they also have higher relative velocity fluctuations as seen in Figure 2 (high σ_r). Of particular interest is the identification of slow flow regions that are not associated with changes in the magnetic polarity of the flow. These are now generally accepted to be "pseudostreamers" that come from streamer-like flows at the Sun but with bipolar regions below the open field lines, leading to the unipolar flow. (See

Table 1
Mean Values (Centroids) of the Eight Quantities Used in the Analysis (Columns) for Each of Eight Types of Wind (Rows)

Color	V_p	n_p	σ_c	o7to6	σ_r	fetoo	aveqfe	b_0	N_k/N
Red (CH+)	504.1889	4.8210	0.8140	0.1684	−0.2705	0.1576	10.1349	7.5191	0.246
Blue (NCH−)	406.2666	7.5016	−0.4670	0.2762	−0.6343	0.1881	10.0653	6.8543	0.113
Green (ICME)	464.4140	8.2996	0.0950	0.5480	−0.6402	0.2215	11.5302	10.4567	0.095
Purple (NCH+)	394.6382	6.6545	0.5520	0.2840	−0.5866	0.2010	9.9442	6.8353	0.154
Cyan (NCH−)	475.5256	5.4553	−0.6446	0.4547	−0.4425	0.1667	11.9122	6.8356	0.043
Magenta (PS)	477.4309	5.4904	−0.8311	0.1415	−0.2954	0.1507	9.9792	7.1535	0.150
Black (CH−)	682.3386	3.3511	−0.8070	0.0665	−0.2579	0.1052	10.4443	6.9685	0.146
Orange (IR)	451.6607	11.3938	−0.1592	0.2711	0.0484	0.1829	10.3147	9.9361	0.052

Note. The categories are CH, coronal hole; NCH, noncoronal hole; ICME; PS, pseudostreamer; and IR, interaction region; with + and − indicating the magnetic sector. N_k/N is the fraction of the number of points in the cluster.

Xu & Borovsky 2015 and references therein.) Here these regions appear to be the magenta states. They show slow speed but high cross-helicity. This is consistent with a careful examination of the results of Ko et al. (2018), where row 3, column 1 of their Figure 5 has systematically higher green points (pseudostreamers) than black points (current sheet crossings).

A summary of the expert identifications of regions by cluster is given in the first column of Table 1. The values of the centroids of the clusters shown there are consistent with the verbal descriptions above, e.g., ICMEs have low cross-helicity, high mean field (b_0), and high values of average iron charge state (aveqfe), as expected. Coronal holes are characterized by generally higher speeds (v_p) and cross-helicity. Pseudostreamers are characterized by high cross-helicity but low speed, and interaction regions have the expected high density resulting from plasma compression. All these characteristics are consistent with the discussion above.

3.2. Clusters without Charge State Information

To see whether it is possible to identify the similar sets of solar wind regions without the use of composition variables, we use a set of variables developed by Xu & Borovsky (2015) in their analytic formulation of the state determinations. The use of $T_{\rm exp} = 1.2 \times 10^4 (V_p/235.0)^3$ (the "expected" temperature for a given speed of wind, V_p) gives one variable as the ratio of this quantity to the measured proton temperature: $T_{\rm ratio} = T_{\rm exp}/T_p$. A second variable is the Alfvén speed $V_a = |b|$ (in "Alfvén speed units" as in Equation (4)). The third variable is the entropy $S = T_p/n^{2/3}$. It is necessary to take the log of all values, as done by Xu and Borovosky, to spread out the somewhat clumped distribution of low values of the quantities. This kind of scaling is common in k-means analysis. It represents a strength and a weakness of the method in that it gives great flexibility in the description, but puts an increased burden on the verification process. In our first use of these variables, we choose four clusters to see how results compare to the study of Xu and Borovsky. Figure 3, which is in the same format as Figure 1, shows that the variable set is a good proxy (as expected) for the composition-based variable set. The four colors nicely correspond to the Xu–Borovsky states: purple is the coronal hole wind, red the current sheet wind, blue the streamer belt (and related pseudostreamer wind), and green represents the transients (ICMEs). These identifications are further confirmed by the values in Table 2, in which, for example, the ICME row shows low cross-helicity, high iron charge states, and high mean field (B_0), and coronal hole flow is typically fast with low o7to6 and high cross-helicity. There are some remarkable agreements with the case above. For example, the first ICME, in May, is split as above, and the possible ICME in early July is not found here or above. Interestingly, the weakly identified ICME in mid-August (the first of a possible pair) is more strongly represented here. The two cases in July that are not found manually are found both here and above. The green labeled state has an enhanced average charge state of iron and o7to6 ratio (as also above), as well as large magnetic field magnitude. Thus the new set of variables captures charge state information without the need for a detailed composition instrument, which is seldom included on spacecraft. Further evidence of the efficacy of k-means is shown in Figure 4, which should be compared to Figure 3 of Xu & Borovsky (2015), which shows a similar structure although the present case is not based on pre-chosen events.

As an example of the importance of the choice of state variables, Figure 5 shows what happens if the temperature variable is omitted from the set of three. In this case, the plot of Alfvén speed versus entropy gives a full characterization of the resulting clusters, and all the algorithm can do is make four more-or-less equal groups. This constitutes a test that the algorithm is working, but it does not give a useful clustering.

3.3. Further Comparison of the Two Cases Above

To make a more direct comparison with the first case above, the variable σ_c, not in the Xu–Borovsky list, but also independent of composition, is added to the list to keep track of magnetic sectors. For what is shown here, k is taken to be 8 for agreement with the first case above. In Figures 6 and 7, orange and blue represent the coronal hole wind in the two sectors, black shows ICMEs, red captures pseudostreamers, and green and cyan the noncoronal hole wind. Although density was not included, interaction regions appear as magenta and purple (depending on sector), but they are not well distinguished from other noncoronal hole wind. The split of the ICME in May, the nonoccurrence of an ICME in early July, the possible new ICMEs in July, and other features are common to this case and the first. There are details that differ (e.g., still a stronger first member of the possible pair of ICMEs in mid-August); these cases will guide future work.

There are various statistical methods used to determine how "good" the clustering is, as well as to attempt to determine the "best" value of k. The "elbow" method finds J for k from 1 to past the point where little change occurs, and typically uses the k value where the decrease in J becomes less dramatic as the best value. (The value of J goes to zero as $k \rightarrow N$.) This is a

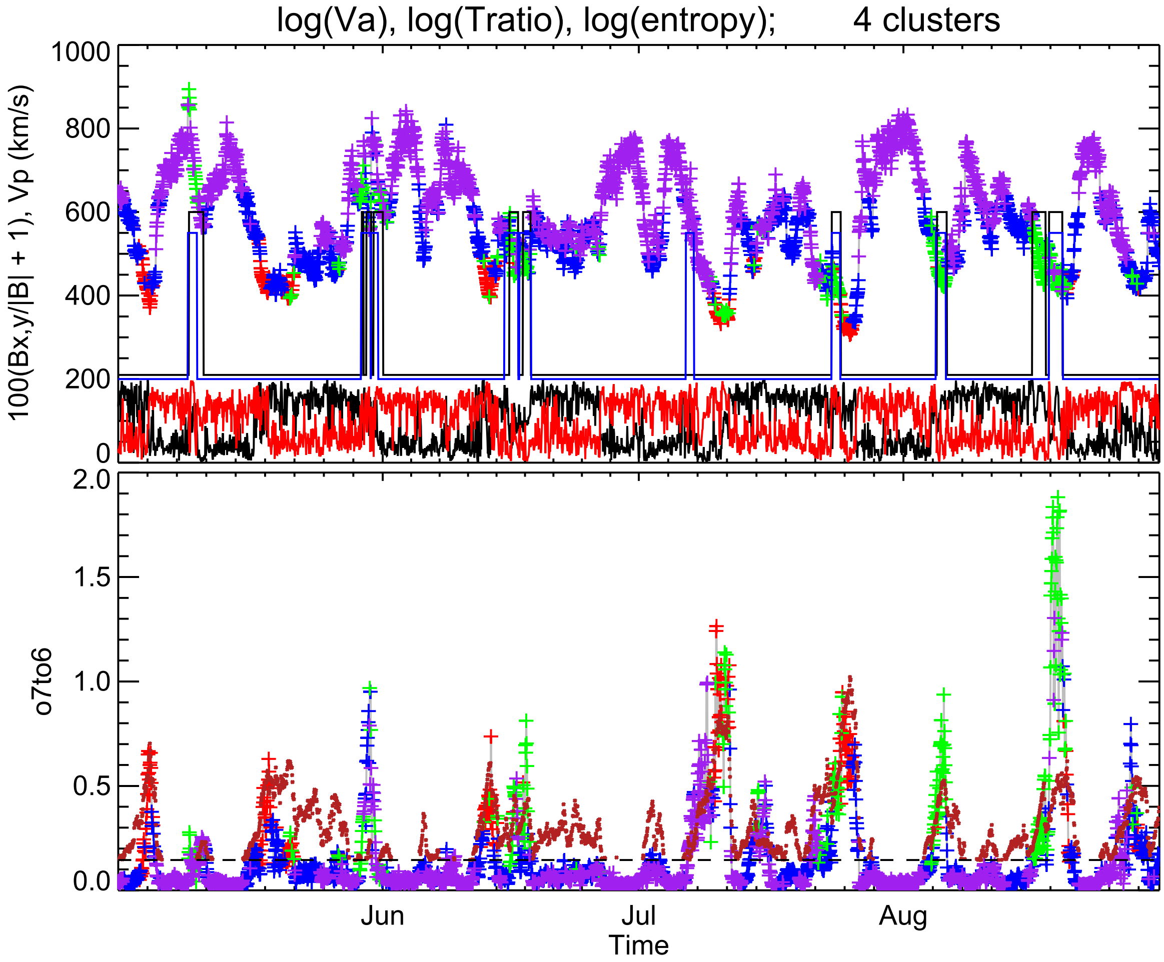

Figure 3. A $k = 4$ set of clusters for four months of 2003 using the variables defined by Xu and Borovsky. The panels are as in Figure 1.

Table 2

Means of Eight Properties (Columns) of Four Types of Wind (Rows) found with the Three Variables of Xu & Borovsky (2015)

Color	V_p	n_p	$\|\sigma_c\|$	o7to6	σ_r	fetoo	aveqfe	b_0	N_k/N
Red (SR)	378.4190	9.6784	0.5074	0.3325	−0.5070	0.2001	10.2638	6.4704	0.215
Blue (PS/SB)	475.9861	5.6409	0.6919	0.1659	−0.3734	0.1532	10.1628	7.4252	0.449
Green (ICME)	458.9792	6.6945	0.4618	0.6466	−0.4685	0.2550	11.4062	9.1196	0.083.
Purple (CH)	622.5383	3.2403	0.7157	0.1304	−0.2853	0.1333	10.3572	8.2802	0.253

Note. The categories identified are SR, sector reversal; PS/SB, pseudostreamer/streamer-belt; ICME; and CH, coronal hole. Note that the absolute value of the cross-helicity is used, since the two sectors cannot be distinguished by the three variables used here.

highly subjective criterion, and, at least in the case here, it adds little to the discussion. Figure 8 shows the plot of J versus k corresponding to the set of eight variables above. A case could be made for a "best" k of 3 or 4, whereas it is clear from the above discussion that meaningful distinctions are apparent up to at least $k = 8$. Even the $k = 10$ case introduces a clearly identifiable new category of very fast ICMEs (not shown in the figures above). Whether or not this is a distinct category in terms of its origin at the Sun is not clear, but the method does suggest a possibly important distinction. Deeper insight will require detailed examination of each case using other solar imaging and in situ data.

The "silhouette" method measures the compactness of each cluster compared to the distance to the next closest cluster. For k-means, the clusters are, by definition, distinct and non-overlapping, so all this test shows is how compact the clusters are. From the σ_c–σ_r plots and other pairs we find that the clusters do not present themselves as being highly compact. This implies that the edges of the distributions are somewhat fuzzy, perhaps due to the evolution of the plasma from the Sun. We can hope that measurements from *Parker Solar Probe* (*PSP*) nearer the Sun will yield more compact clusters and thus clearer states, but the physics-based criteria used here imply that the clusters found are meaningful.

4. Conclusions

The above results illustrate the likely utility of unsupervised machine learning techniques in general, and the k-means method in particular, for identification of distinct states of the solar wind. The method is computationally fast and only involves subjective decisions in the choice of state variables, not in any subsequent decisions on how to categorize wind

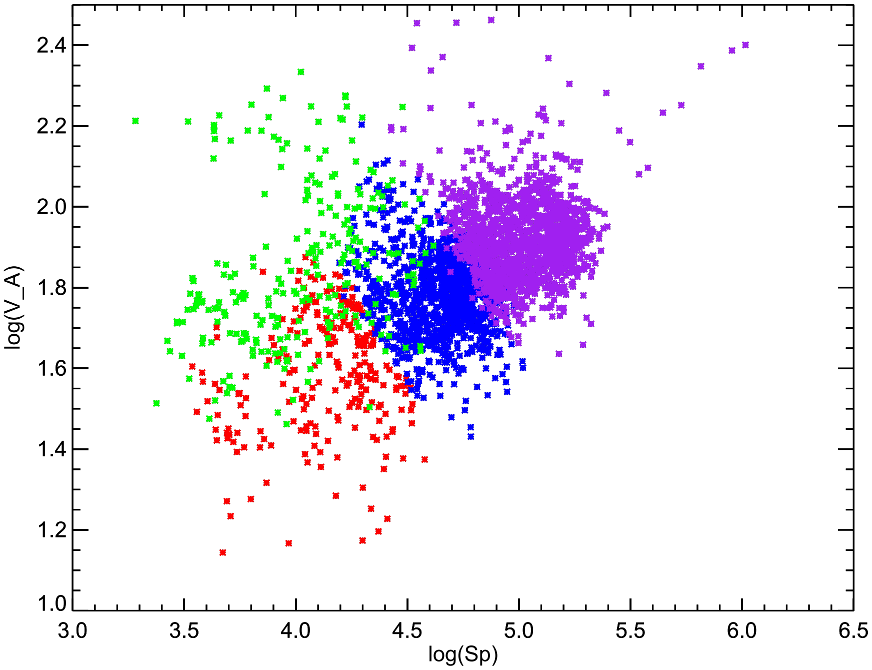

Figure 4. Clusters in the previous figure, plotted in the $V_A - S_p$ space; see Figure 3 of Xu & Borovsky (2015).

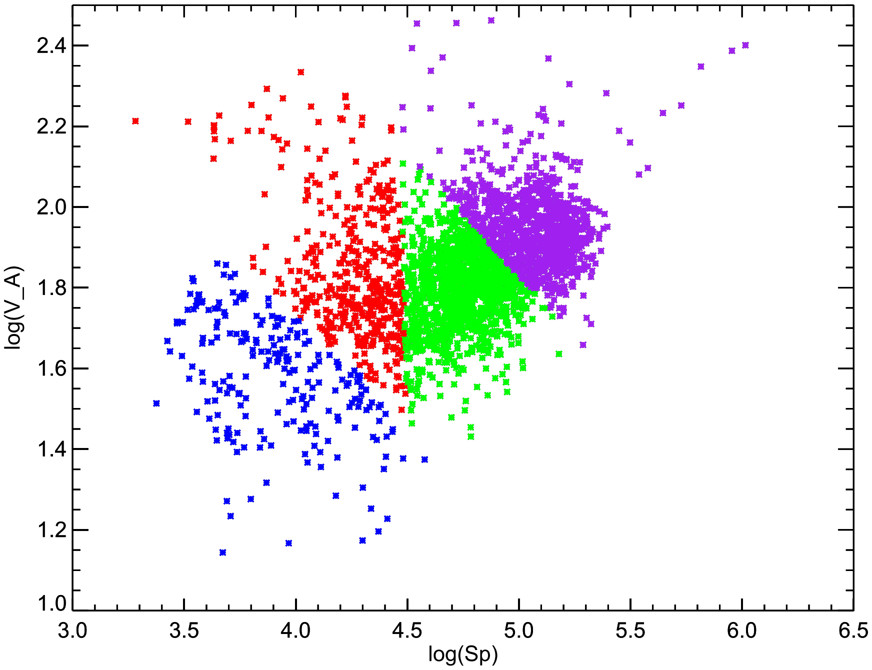

Figure 5. Clusters for two variables, S_p and V_A, and $k = 4$. Note that all the algorithm can do in this case is make four similar partitions in the 2D space, unlike the divisions in the previous figure where a third variable allows more physically significant distinctions.

states. The ability to vary the number of clusters provides a way of systematically looking for substructures. The natural next steps are to decipher the meaning of the various complex wind regions and the cases of discrepancies that remain between various methods. More quantitative tests will be useful to see the degree to which, for example, different sets of variables identify the same states.

There are possible complications that none of the current solar wind classification methods address. For example, it may be that the solar wind consists of distinct as well as mixed/hybrid states. This possibility should be considered and can be exposed to some extent using the clustering techniques. There are already suggestions of this in the solar wind. For example, ICMEs found here are often mixed with other states in time, but there may be deeper meanings to this.

The standard clustering methods such as *k*-means are based on the underlying assumption that each point can be treated independently and is not causally connected with its neighboring points. To partially account for this, we explicitly introduced temporal correlations through the use of cross-helicity and residual energy as variables. (Surprisingly, an interesting set of correlations arises in the "σ_c–σ_r space.")

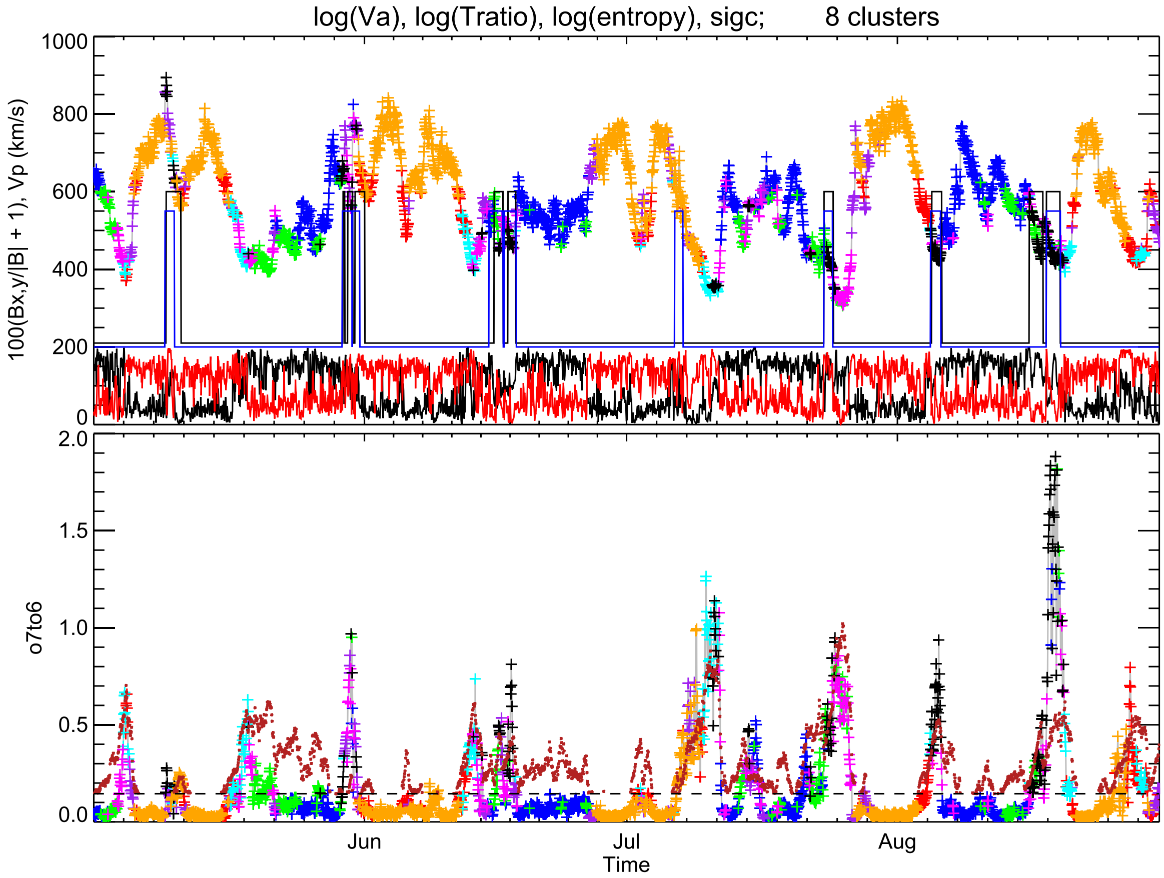

Figure 6. A $k = 8$ set of clusters for four months of 2003 using the variables defined by Xu and Borovsky along with σ_c. The panels are as in Figure 1. The clusters are identifiable as CH– (orange), CH+ (blue), ICMEs (black), PS/SB– (red), PS/SB+ (green), NCH– (cyan), IR–(?) (magenta), IR+ (purple).

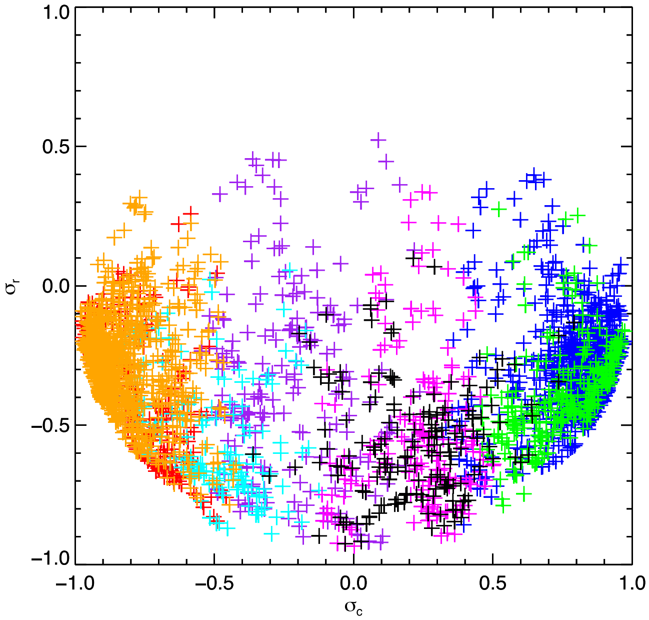

Figure 7. A $k = 8$ set of clusters for the same case as in the previous figure.

Ideally, the unsupervised method should find such temporal relationships on its own. A recently developed unsupervised technique addresses this shortcoming by taking into account any underlying temporal correlations in time series data (Madiraju et al. 2018). Whether this technique will yield better solutions to the categorization of the solar wind remains to be seen. As a prelude to such a study, it would be interesting to examine the average lifetime of the different states derived from clustering.

Another deficiency of the standard clustering techniques is that they do not include learning of the feature space; the state variables must be specified by the user. The proper choice of feature space can have a dramatic impact on the clustering results, as illustrated by Figure 5. We are currently working on the development of a deep learning-based clustering technique that includes learning representations, and we will explore its application to the solar wind.

One significant outcome of this study was the further establishment of a set of variables that can serve as a proxy for composition measurements, namely the set for Figures 3 and 4. This was known to some degree before, but the comparisons here confirmed the efficacy of these variables as proxies for composition measurement. The non-composition proxy not only opens up times and regions where composition data are not available, but also provides a straightforward means of performing classifications at higher time resolution. The ease and speed of the k-means method is also conducive to the studies with higher time resolution.

These techniques used here may be helpful in determining the origin of solar wind parcels, and may also be applicable to many other space plasmas. The methods may provide an automated “scientist in the loop” for determining when to download burst mode data from spacecraft. There are a wide range of other possible directions to extend this work, including using other clustering schemes, seeing what different sets of

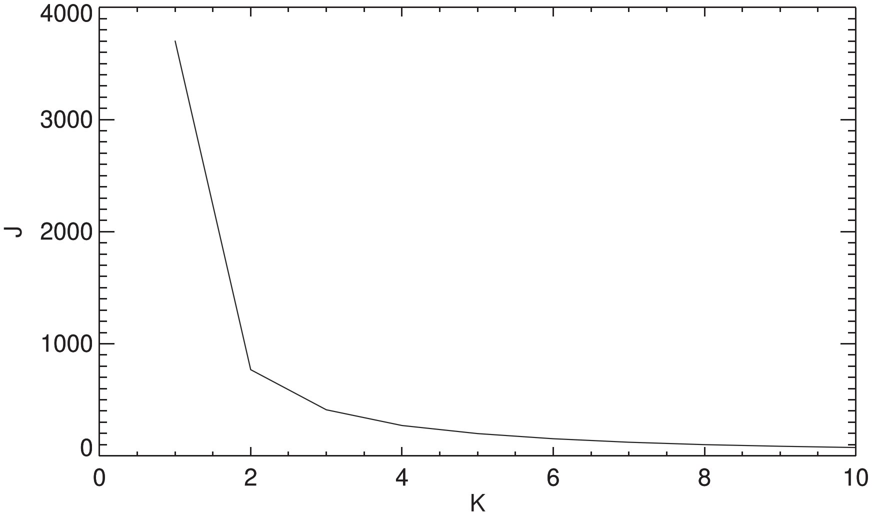

Figure 8. The value of J from Equation (1) as a function of k for the set of variables used in Figure 1.

variables reveal, and using these unsupervised learning states to provide labeled states to train supervised methods, such as neural networks.

Our most immediate next step will be to examine higher resolution data to determine whether the current clusters become more complex or just add more points to each cluster region. If the latter is true, then we can have increasing confidence in the identifications of wind state made by k-means methods. If the picture becomes increasingly complex at higher resolution, or even just in light of existing discrepancies between the cases shown above, we will need to examine the details of these cases both in situ and in correlation with studies of projections back to the Sun. It will also be revealing to see whether the state identification becomes clearer closer to the Sun in *PSP* data, which should help to sort out complexities at the source from those due to propagation.

D.A.R. is supported in this work by Research and Technology Operating Plan (RTOP) funding through the NASA HP Internal Scientist Funding Model. H.K. is supported by Analytics Ventures. T.S. is self-supported for this effort. Y.-K.K. was supported by NNH10AO82I and the Chief of Naval Research. S.T.L. acknowledges funding from NASA grants: NNX16AH01G, NNX17AI94G and NSF grant: AGS1460170. We thank the *ACE* MAG, SWICS, and SWEPAM teams for provision of the data via NASA's CDAWeb. The specific data sets are given by NASA SPASE IDs: spase://VSPO/NumericalData/ACE/SWICS_SWIMS/L2/PT1H, spase://VSPO/NumericalData/ACE/SWEPAM/L2/PT1H, and spase://VSPO/NumericalData/ACE/MAG/L2/PT1H. Entering the ID as a text restriction in the Heliophysics Data Portal at https://heliophysicsdata.gsfc.nasa.gov will bring up the resource and clicking on its name will give access and other information.

Facility: ACE.

ORCID iDs

D. Aaron Roberts https://orcid.org/0000-0001-6565-2921
Yuan-Kuen Ko https://orcid.org/0000-0002-8747-4772
Susan Lepri https://orcid.org/0000-0003-1611-227X

References

Bavassano, B., Pietropaolo, E., & Bruno, R. 1998, JGR, 103, 6521
Borovsky, J. E., Denton, M. H., & Smith, C. W. 2019, JGRA, 124, 2406
Camporeale, E., Carè, A., & Borovsky, J. E. 2017, JGRA, 122, 10910
Jian, L., Russell, C. T., Luhmann, J. G., & Skoug, R. M. 2006, SoPh, 239, 393
Ko, Y.-K., Roberts, D. A., & Lepri, S. T. 2018, ApJ, 864, 139
MacQueen, J. 1967, Proc. of the Fifth Berkeley Symp. on Mathematical Statistics and Probability (Berkeley, CA: Univ. California Press), 281, https://projecteuclid.org/euclid.bsmsp/1200512992
Madiraju, N. S., Sadat, S. M., Fisher, D., & Karimabadi, H. 2018, arXiv:1802.01059
Marsch, E., Rosenbauer, H., Schwenn, R., Muehlhaeuser, K.-H., & Denskat, K. U. 1981, JGR, 86, 9199
McGranaghan, R. M., Bhatt, A., Matsuo, T., et al. 2017, JGRA, 122, 12586
Neugebauer, M., Reisenfeld, D., & Richardson, I. G. 2016, JGRA, 121, 8215
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn., 12, 2825
Richardson, I. G., & Cane, H. V. 2010, SoPh, 264, 189
Roberts, D. A., Goldstein, M. L., Klein, L. W., & Matthaeus, W. H. 1987, JGR, 92, 12023
Wicks, R. T., Roberts, D. A., Mallet, A., et al. 2013, ApJ, 778, 177
Xu, F., & Borovsky, J. E. 2015, JGRA, 120, 70
Zhao, L., Landi, E., Lepri, S. T., et al. 2017, ApJ, 846, 135
Zhao, L., Zurbuchen, T. H., & Fisk, L. A. 2009, GeoRL, 36, L14104